Introduction to
BIG DATA
and
MACHINE LEARNING
for Survey Researchers

Trent D. Buskirk, Ph.D.
Novak Family Professor of Data Science
Bowling Green State University

Webinar sponsored by the ASA Government Statistics Section
in partnership with the ASA Social Statistics Section
September 19, 2019

BGSU®

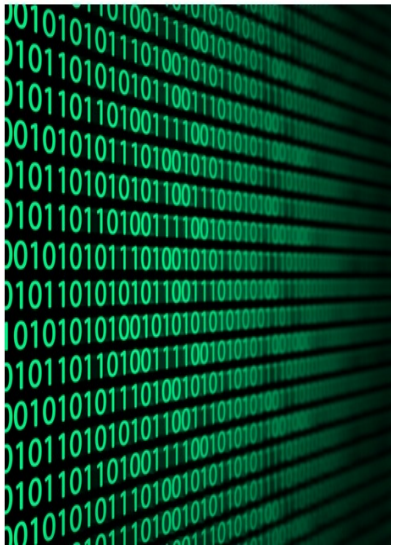
Outline for Today's Webinar

- Overview of Big Data
  - What is Big Data?
  - Why does it matter?
  - What's in this for Survey Researchers?
  - Big Data Panacea??
- What is Machine Learning?
  - What is machine learning?
    - **Types of learners**
  - Classical Statistical Approaches versus Statistical Machine Learning
  - Model Evaluation/Validation

Outline, Continued

- Basic Introduction to common Machine Learning Algorithms
  - *K-Means Clustering*
  - *k-nearest neighbors*
  - *Classification and Regression Trees*
  - *Random Forests*
  - *Extra Trees*

- *Resources for Machine Learning*
  - *Training*
  - *R meta packages for ML*
  - *Other Open Source and Proprietary Resources*



Brief Overview of Big Data

What is Big Data?

Why does it matter?

What's in this for Survey Researchers?

Opportunities and Challenges…

What is Big Data?

Big data describes the collection of complex and large data sets such that it's difficult to ***capture, process, store, search and analyze*** *using conventional data base systems. Its uses are shaping the world around us, offering more qualitative* ***insights*** *into our everyday lives.*

Ben Walker, **Marketing Executive** at vouchercloud**, 2016**

http://www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillion-bytes-of-data-created-daily/

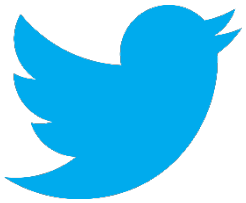
The BIG data revolution…

"Whilst there may be a 'big data revolution' underway, it is not the size or quantity of these data that is revolutionary.

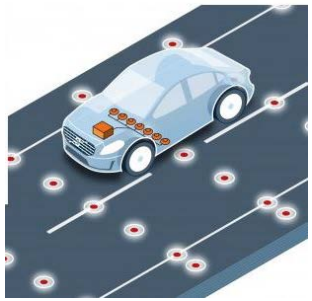
The revolution centers on the increased availability of new types of data which have not previously been available for social research."

R. Connelly and Colleagues, 2016
http://bit.ly/2k0V7GM

Today's Big Data Sources...



BIG DATA



‘Big Data’ – you already know

- Paradata (ideally)
  - real time, longitudinal, “everybody”
- Useful for survey researchers to monitor
  - inactive interviewer
  - coverage composition
- Allows action while it happens to save cost
- Even more opportunity in web surveys
  - mouse movements
  - response times
  - missing data
- And more opportunities for triangulating various data sources together…

Surveys and Beyond: Triangulation

- 23andMe: https://www.23andme.com/research/

Be part of something bigger.

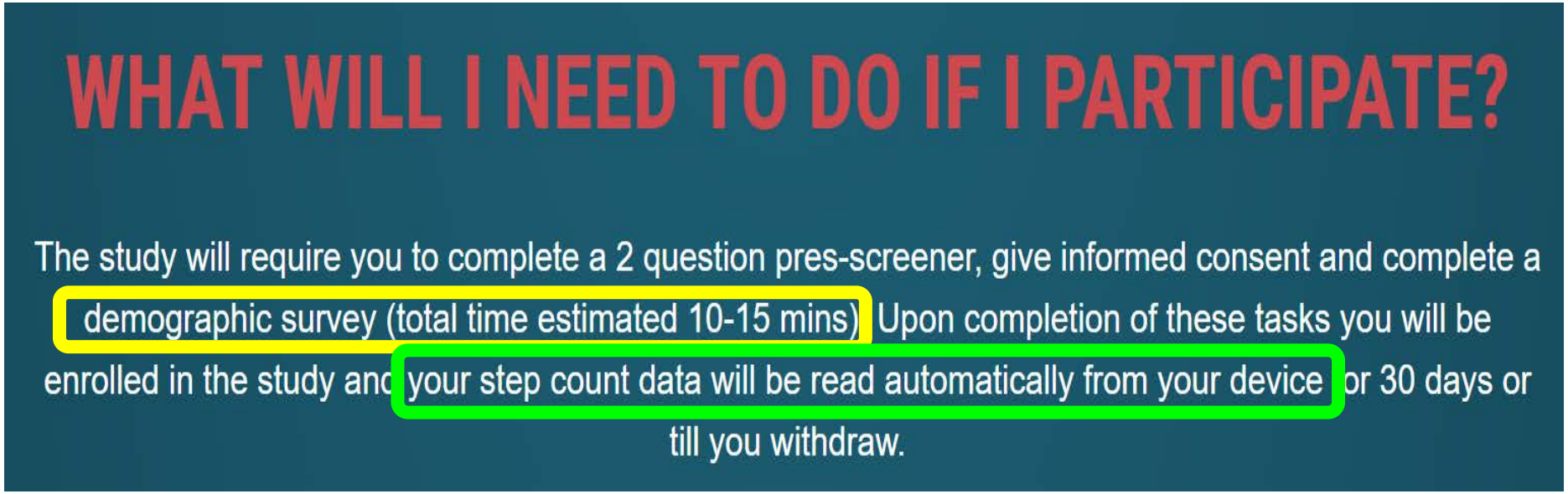
You can make a difference by participating in a new kind of research—online, from anywhere.

As a customer, you can answer online survey questions, which researchers can link to your genetic data to study topics from ancestry to traits to disease. Your contribution helps drive scientific discoveries.

You can always choose to opt into or out of research.

Active + Passive Survey Data Collection

- http://trialx.com/americawalksstudy/



Administrative Data

- By far one of the most common forms of big data used in the social sciences relates to administrative data.
- Administrative data are a type of "found data"
  - may not have been collected for research purposes;
  - may be messy
  - but usually represent a known population/sample.

ELSEVIER

Social Science Research

Volume 59, September 2016, Pages 1–12

Special issue on Big Data in the Social Sciences

The role of administrative data in the big data revolution in social science research

Roxanne Connelly[a], Christopher J. Playford[b], Vernon Gayle[c], Chris Dibben[d]

Show more

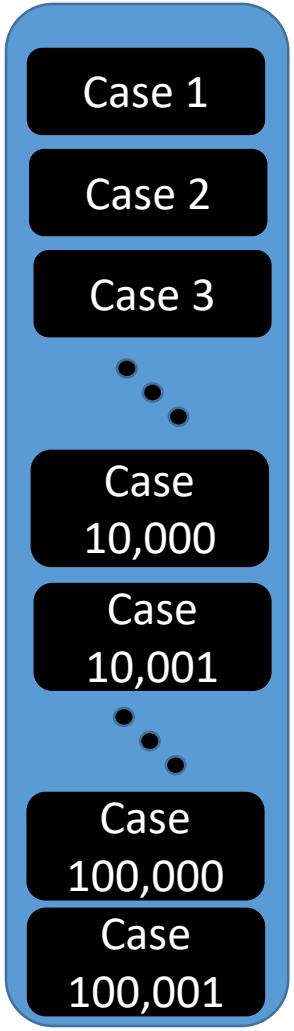
http://dx.doi.org/10.1016/j.ssresearch.2016.04.015

Get rights and content

Big Data – Long or Wide?

- A natural perspective within the Big Data paradigm is to think about Big Data in terms of the sheer number of CASES available.

- But in social science and education applications another perspective might not be how long, but rather how wide a data set can become.



Data and Social Scientists and the era of Big Data

"Big Data provides inexpensive and timesaving means to tap unchartered sources of empirical evidence."

A. Tokhi and C. Rauh, Berlin Social Science Center

"Data Scientists have significantly more experience with large datasets but they tend to have little training in how to infer causal effects in the face of substantial selection. Social scientists must have an integral role in this collaboration; merely being able to apply statistical techniques to massive datasets is insufficient. Rather, the expertise from a field that has handled observational data for many years is required."

J. Grimmer, Stanford University

Bias resulting in coverage has become big news!

Artificial Intelligence Can Reinforce Bias, Cloud Giants Announce Tools For AI Fairness

Paul Teich Contributor ⓘ

Enterprise & Cloud

I write about new technologies and usage models transforming business.

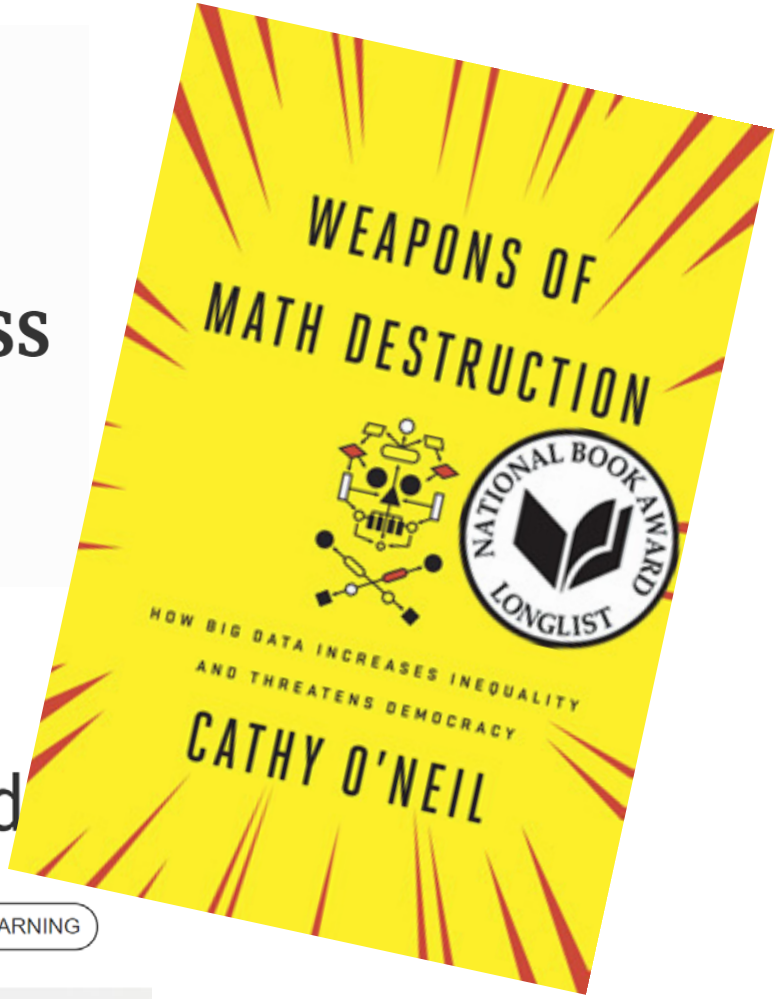
Google's New Machine Learning Curriculum Aims to Stop Bias Cold

Nate Swanner

October 24, 2018

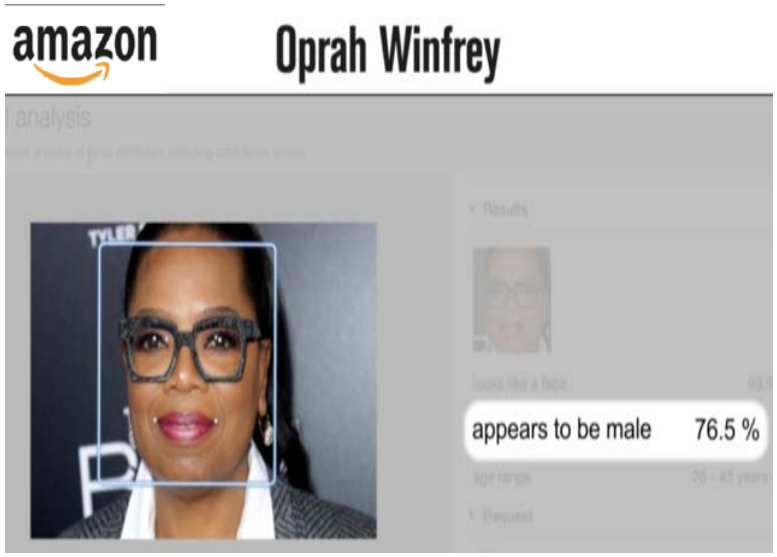
3 min read

ARTIFICIAL INTELLIGENCE | GOOGLE | MACHINE LEARNING



More Examples of Bias in Algorithms

- In a recent Time article, "Artificial Intelligence Has a Problem With Gender and Racial Bias. Here's How to Solve It," Joy Buolamwini (2019) report how facial recognition software used by Amazon, Microsoft and other large tech firms have little misclassification of gender for white users but this error soars to above 35% for darker skinned women, for example.
  - http://time.com/5520558/artificial-intelligence-racial-gender-bias/



Total Error Frameworks...

- Tracking error in survey data has long been investigated and implemented in practice under the Total Survey Error framework.
- Biemer and Amaya (2018) extend this framework to the big data context by proposing the Total Error Context.
- Briefly, the total error is the sum of errors made in the columns of the data set as well as errors made in the rows of the data frame.
  - Data Encoding Errors (Column related Errors related to columns)
    - **Specification, Measurement Error, Data Processing Errors**
  - Sample Recruitment Errors (Errors associated with rows)
    - **Coverage, Nonresponse, Self-Selection**

Biemer and Amaya (2018): https://www.bigsurv18.org/program2018?sess=18#91

Credits and a Plug

Why Machines Matter for Survey and Social Science Researchers:

Exploring Applications of Machine Learning Methods to Design, Data Collection and Analysis

Trent D. Buskirk and Antje Kirchner, 2019

Exploring New Statistical Frontiers at the Intersection of Survey Science and Big Data: "BigSurv18" Conference www.bigsurv18.org

A new landscape for survey research

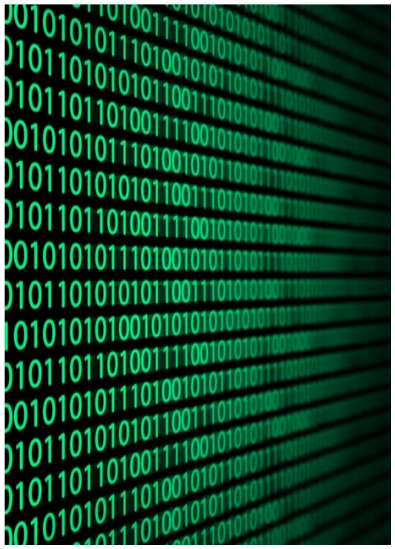
- Buskirk and Kirchner (2019) posit a new survey research landscape in which we are:
  - Reimagining traditional survey research by leveraging new machine learning methods that improve efficiencies of traditional survey data collection, processing, and analysis;
  - augmenting traditional survey data with non-survey data (administrative, social media, or other Big Data sources) to improve estimates of public opinion and official statistics;
  - enhancing official statistics or estimates of public opinion derived from big data or other non-survey data;

New Survey Research Landscape

- And a landscape in which we as survey researchers are:
  - comparing estimates of public opinion and official statistics derived from survey data sources to those generated from Big Data or other non-survey data exclusively;
  - exploring new methods for enhancing survey and non-survey data collection and gathering, processing and analysis;
  - adapting and modifying current methods for use with new data sources and developing new techniques suitable for design and model-based inference with these data sources;
  - contributing survey data, methods and techniques to the Big Data ecosystem.

Leveraging the Power of Big Data

- So how can we leverage the power and potential of big data within this new landscape for survey research and social science?
- Limitations of current statistical methods exist when faced with large amounts of records and increasing numbers of potentially correlated variables.
- Machine learning methods offer flexible approaches for adapting to the complexity of big data sources in terms of both the number of cases and the number of variables.



What is Machine Learning?

What is Big Data?

Why does it matter?

What's in this for Survey Researchers?

Opportunities and Challenges…

Machine Learning

- Wikipedia defines:

Machine learning (ML) is the scientific study of **algorithms** and **statistical models** that computer systems use to effectively perform a specific task **without using explicit instructions**, relying on patterns and inference instead. It is seen as a subset of artificial intelligence.

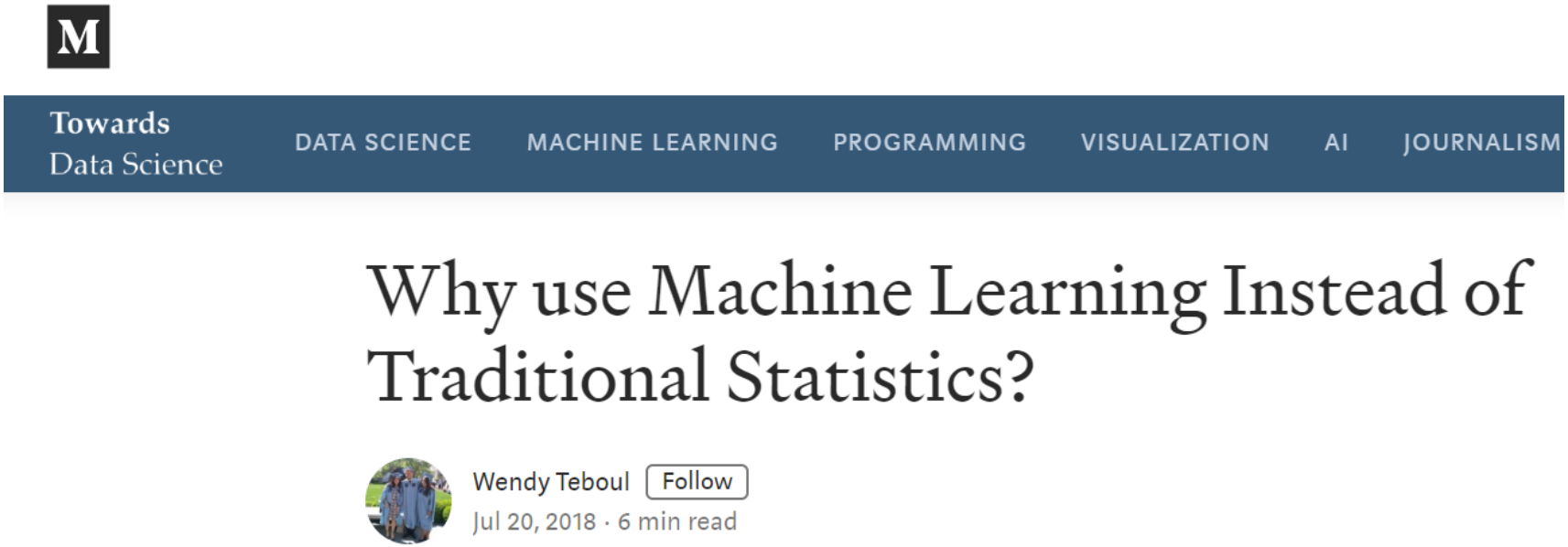
Machine learning algorithms build a **mathematical model of sample data**, known as "**training data**", in order to make **predictions or decisions** without being explicitly programmed to perform the task.

The Machine Learning Context

- Emphasis moves away from statistical inference, per se, to statistical prediction.
  - Applications of machine learning methods often focus more on accurate predictions ("who/which") rather than on inference (or on understanding the why).
  - Machine learning models are cross validated with key metrics focusing on prediction or classification accuracy, area under the curve, true positive and true negative rates and the like; rather than on p-values and the significance of predictors.

Machine Learning versus Traditional Statistics?

- https://towardsdatascience.com/why-use-machine-learning-instead-of-traditional-statistics-334c2213700a



Differentiating Aspects of ML…

- "We see that the goal of ML is not to come up with knowledge about the data ('this is the real phenomenon, this is how it works') but rather with a workable and reproducible model for which the error tolerance is determined by the project we are undertaking."
- "Learning methods are in fact necessary to deal with plenty of problems. They ignore our lack of knowledge about the data by not prompting us to choose a model."

- Wendy Teboul

ML Differentiated…

- Machine learning models provide various degrees of interpretability, from the highly interpretable lasso regression to impenetrable neural networks, but they generally sacrifice interpretability for predictive power.
- *"The major difference between machine learning and statistics is their purpose. Machine learning models are designed to make the most accurate predictions possible. Statistical models are designed for inference about the relationships between variables."*

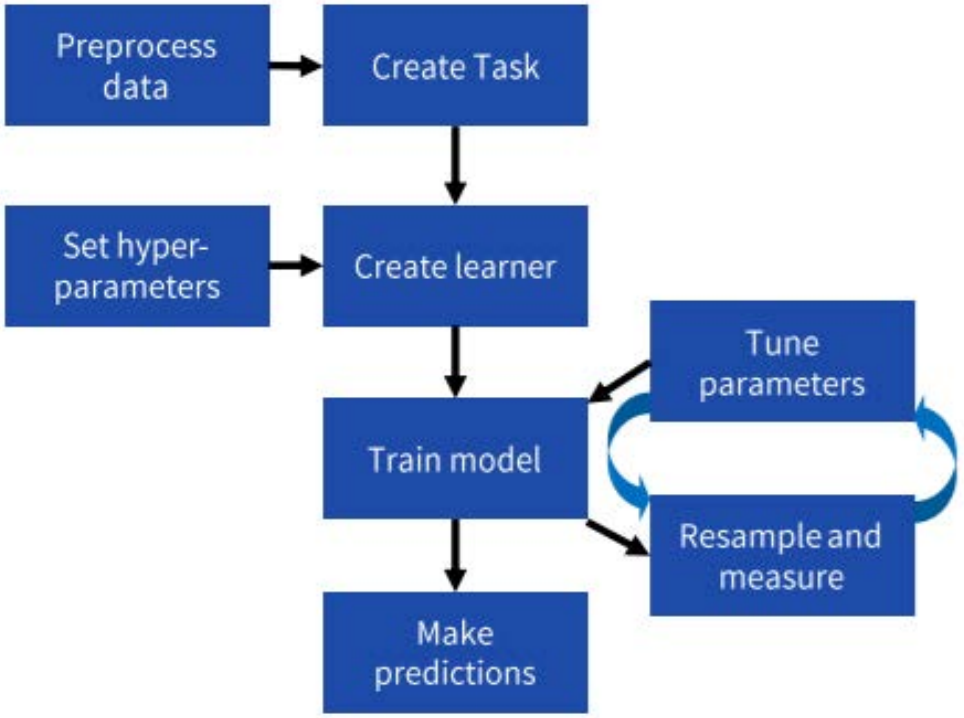
  Matthew Stewart, http://bit.ly/MatthewStewart

Prediction versus Explanation

- Machine learning algorithms can also describe relationships between predictors and outcomes
  - But many times, this is not the end goal.
  - End goal is to develop well tuned predictive "models" that can be applied to new data.
  - Here models are evaluated based on their *predictive power.*

- *An excellent overview of the model building process within the context of prediction is given by Shmueli (2010).*
  - *See: http://bit.ly/1F27hGT*

Machine Learning Workflow



- *Machine learning models have so called tuning or hyperparameters that function to improve prediction akin to nuisance parameters in parametric statistics.*

Overview of Machine Learning

- Two main categories of machine learning methods
  - Unsupervised methods: Focused on grouping or segmenting observations based only on covariates (no Ys, only Xs)
    - **Principal Components Analysis**
    - **K-means and K-nearest neighbors**
    - **Hierarchical Clustering**

Think of unsupervised learning as a mathematical version of making "birds of a feather flock together."

Unsupervised Learning Example

- Survey Researchers can use para-data obtained from a survey panel (probability or non-probability) to understand whether there are groups of panelists that are similar on certain survey attributes including:
  - **Survey length**
  - **Incentive amount**
  - **Field period length**
  - **Number of Invitations required before response**
  - **Other attributes**

- The idea would be to determine if there is a group of panelists who consistently take longer surveys, for more money versus those who take shorter surveys for less money, for example.

Unsupervised Learning in the Survey Context

- Generally, a mix of both unsupervised and supervised machine learning approaches can be used for a given study:
  - Auxiliary information could be used to form clusters that were optimal (most heterogeneous) and then survey results could be obtained by incorporating the sampling design (Elliot, 2011: http://1.usa.gov/1OUlyJd)
  - Tailored survey design to predict survey response as a function of hundreds of geographic auxiliary data appended to an ABS sample – use Principal Components to "reduce" the variables into a smaller set of independent signals to be used in a logistic regression model (See Buskirk, West and Burks, 2013).
  - Using geographic variables to predict presence of phone number for a sampling frame of addresses – Olson and Buskirk (2015) first applied factor analysis to create smaller set of factors used for prediction (See Olson and Buskirk, 2015).

Overview of Machine Learning

- Two main categories of machine learning methods
  - Supervised methods: Focused on prediction outcome(s) from a set of covariates (here both Xs and Ys are used).
    - **Logistic/Linear Regression**
    - **Classification and Regression Trees**
    - **Neural Networks**
    - **Support Vector Machines**
    - **Random Forests (many varieties)**
- Broadly speaking, if the outcome variable for a supervised learning algorithm is continuous then you generally have a "regression problem" and if the outcome is categorical, you have a "classification" problem…

Supervised Learning and Survey Data

- Smartphone surveys have begun capturing photo data from respondents (Michaud, Buskirk and Saunders and Nielsen, 2012)
- Suppose a survey asked respondents about their "snacking habits" and asked them to snap a photo of the snack they are currently working on
  - Snacks could be classified as edible or drinkable.
  - Coders could mark photos as edible or drinkable and then the photo data captured from surveys could be classified into one of these categories using the machine learning model that was generated from a training/pilot set of images.

Stanford team creates computer vision algorithm that can describe photos (2014)
http://news.stanford.edu/news/2014/november/computer-vision-algorithm-111814.html

Metrics and Approaches Used to Evaluate Predictions from Machine Learning Methods

- http://bit.ly/BuskirketalSurveyPractice2018

Important Components of the Confusion Matrix

- Positive (P) : Observation is positive (i.e. respondent).
- Negative (N) : Observation is not positive (i.e. nonrespondent).
- True Positive (TP) : Observation is positive, and is predicted to be positive.
- False Negative (FN) : Observation is positive, but is predicted negative.
- True Negative (TN) : Observation is negative, and is predicted to be negative.
- False Positive (FP) : Observation is negative, but is predicted positive.

Example of Confusion Matrix Calculations

- Confusion matrix = cross tabulation of Predicted Outcome (binary) versus Actual Outcome (binary)

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

TP = True positive; FN = False negative

FP = False positive; TN = True negative

confusionMatrix function in caret package in R

```
>require(caret)
>confusionMatrix(predicted_values, actual_values, positive=c("1"))
Confusion Matrix and Statistics
          Reference
Prediction  0  1
         0 52 17
         1  7 24
               Accuracy : 0.76
                 95% CI : (0.6643, 0.8398)
    No Information Rate : 0.59
    P-Value [Acc > NIR] : 0.0002746
                  Kappa : 0.4848
 Mcnemar's Test P-Value : 0.0661926
            Sensitivity : 0.5854
            Specificity : 0.8814
         Pos Pred Value : 0.7742
         Neg Pred Value : 0.7536
             Prevalence : 0.4100
         Detection Rate : 0.2400
   Detection Prevalence : 0.3100
      Balanced Accuracy : 0.7334
       'Positive' Class : 1
```

Confusion Matrix

Accuracy Measures and Tests

Accuracy Measures and Tests

Positive Class Used in Metrics

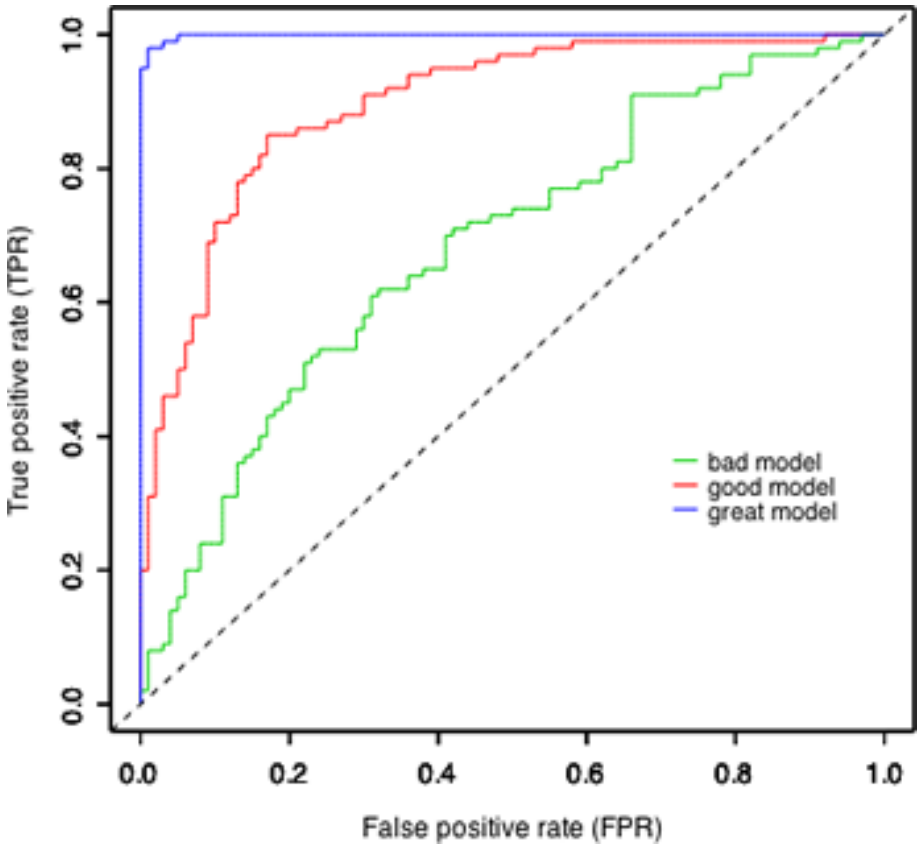
Common Evaluation Metrics Derived from the Confusion Matrix

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	$\frac{TP + TN}{TP + TN + FP + FN}$

ROC Curves…

- AUC ranges from 0 to 1; AUC of random classifier with no information is 0.5 (assuming data are balanced)

SENSITIVITY



1-SPECIFICITY

Introduction to Machine Learning Algorithms

Popular machine learning techniques

- K Nearest Neighbors
- CARTS
- Random Forests
- Extra Trees

Method Overview: K-nearest Neighbors

- For what problems can you apply this method?
  - Regression/Prediction for Continuous Outcome Variables
  - Classification for Categorical Outcome Variables (2 or more levels)
  - Method usually applied when predictors are all continuous/ordinal
- Does this method have any Tuning Parameters?
  - One parameter: k – the number "neighbors"
    - **Here neighbors refer to cases in the data (i.e. rows in the D.S.)**

Method Overview: KNN

How does this Method Work?

Step 1: Determine Neighbors for Each Point in the Data Set

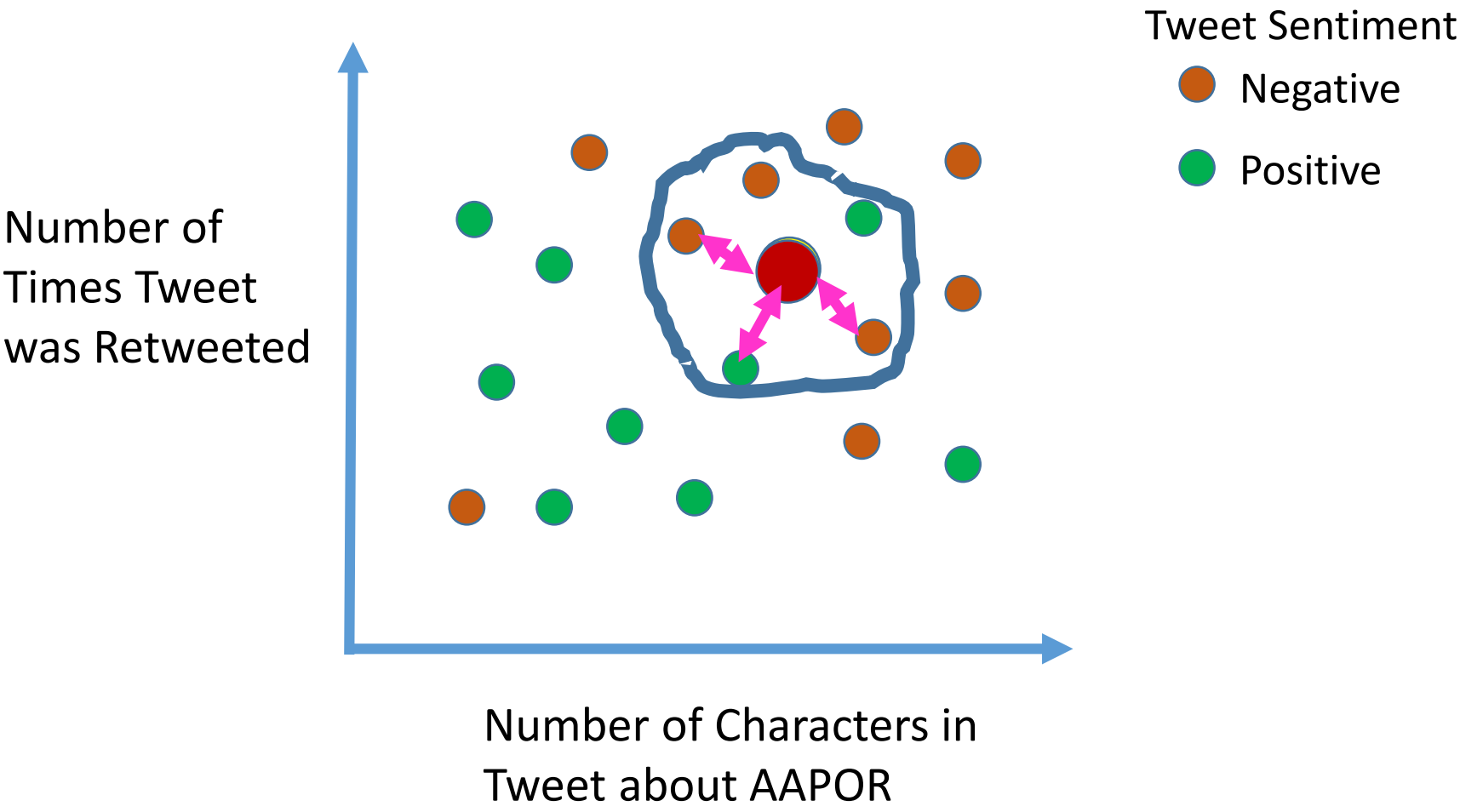
Using a distance measure (Euclidean/City), KNN identifies data points that are “close” in p-dimensional space to each case in the test set for which predicted values are desired.

Step 2: Make the Prediction/Estimate

For classification, the predicted value is the modal class of all k-nearest neighbors in the training set;

For regression, the predicted value is the mean/median of the outcome variable among all k-nearest neighbors in the training set.

Example of KNN with 2 predictors and a Binary Outcome with K=5



Method Overview: KNN

- Insights on Parameters for this Method
  - Small k → low bias, but high variance in estimates;
  - Large k → higher bias but lower variance;
  - For classification applications, k is usually chosen as an odd number (to avoid ties)
  - Other distance functions can also be used, but most common is Euclidean Distance
- Caveats/ Considerations for the Method
  - This method is sensitive to scaling of p-predictors.
  - If all predictors are on different scales/measured in different units, standardizing variables is strongly recommended.

Method Overview: KNN

Caveats/ Considerations for the Method

- This method can be sensitive to the scaling of the p-predictors. If all are on different scales/ measured in different units, standardizing variables is recommended.
  - **This effectively gives every predictor equal weight in determining "similarity" of cases.**
- This method is not recommended for use with nominal categorical predictors without some additional preparation/transformation of these variables.
  - **One alternative is k-mode clustering or providing similarity matrix for each case.**

Method Overview: KNN

- Implementation of this method in R
  - train{caret package; use method="knn");
  - knn{class package};
  - knn.reg{FNN package}
- Examples of this method in the S.R. Literature
  - Multiple Imputation
    - **Jonsson and Wohlin (2004); http://www.wohlin.eu/metrics04.pdf**
  - Image Classification
  - Text Classification
  - Propensity Weighting
  - Binary Prediction from Longitudinal Survey Data
    - **Bryant et al. (1996)**
    - **http://irp.wisc.edu/publications/dps/pdfs/dp109296.pdf**

Example: KNN

- For nearly 20 years (1999-2016) the state of CA has used the Academic Performance Index (API) as a way to measure progress of schools and students within the schools.

- The API datasets are available publically here:
  - https://www.cde.ca.gov/ta/ac/ap/apidatafiles.asp

- These data contain a host of information about each of the schools within the state including:
  - Type of School
  - API scores
  - Aggregated Student Information (ELL, FRL, Mobility, etc.)
  - Aggregated Parent and Teacher Information
  - School Target Indicators

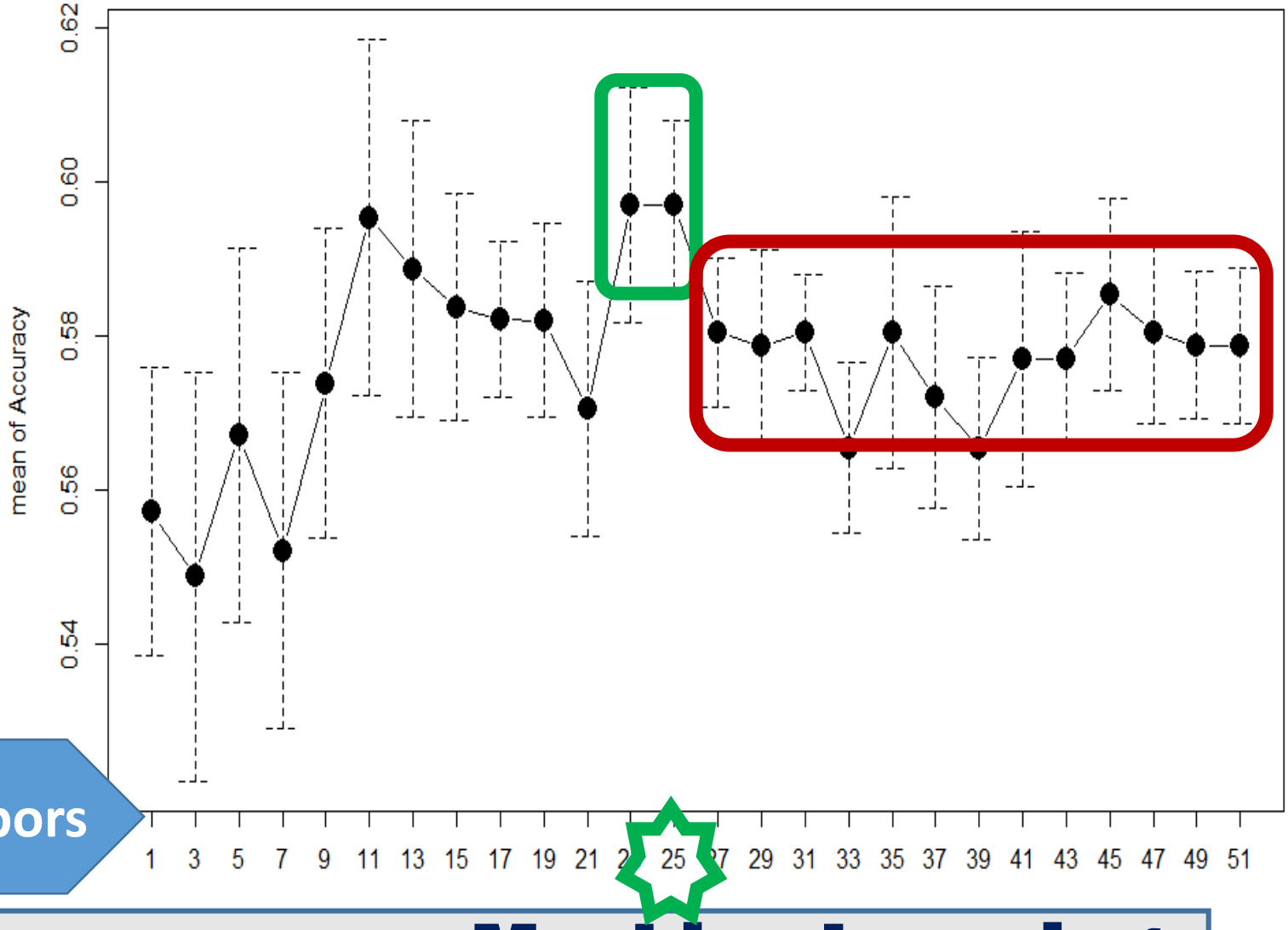
- Data from the 2000 Academic year for the entire population of schools within CA are available in the datafile apipop within the API data files in the survey R package. (data(API))
  - https://cran.r-project.org/web/packages/survey/survey.pdf

KNN – Example, Cont.

- Using these data we want to create a KNN model to predict whether or not a High School has met it's school-wide API growth target for the 2000 academic year using the following predictors:
  - percentage of students eligible for subsidized meals
  - percentage of students who were ELL
  - percentage of students for whom this is the first year at the school
  - percentage of teachers within the school with emergency certification.
- We will create a training sample (80%) and a test sample (20%) using complete cases from 753 HS's; 420 of which met their target (56%) and 333 that did not (44%).
- We will use 5-fold cross validation to determine the optimal value of k – the tuning parameter for the size of the neighborhoods.

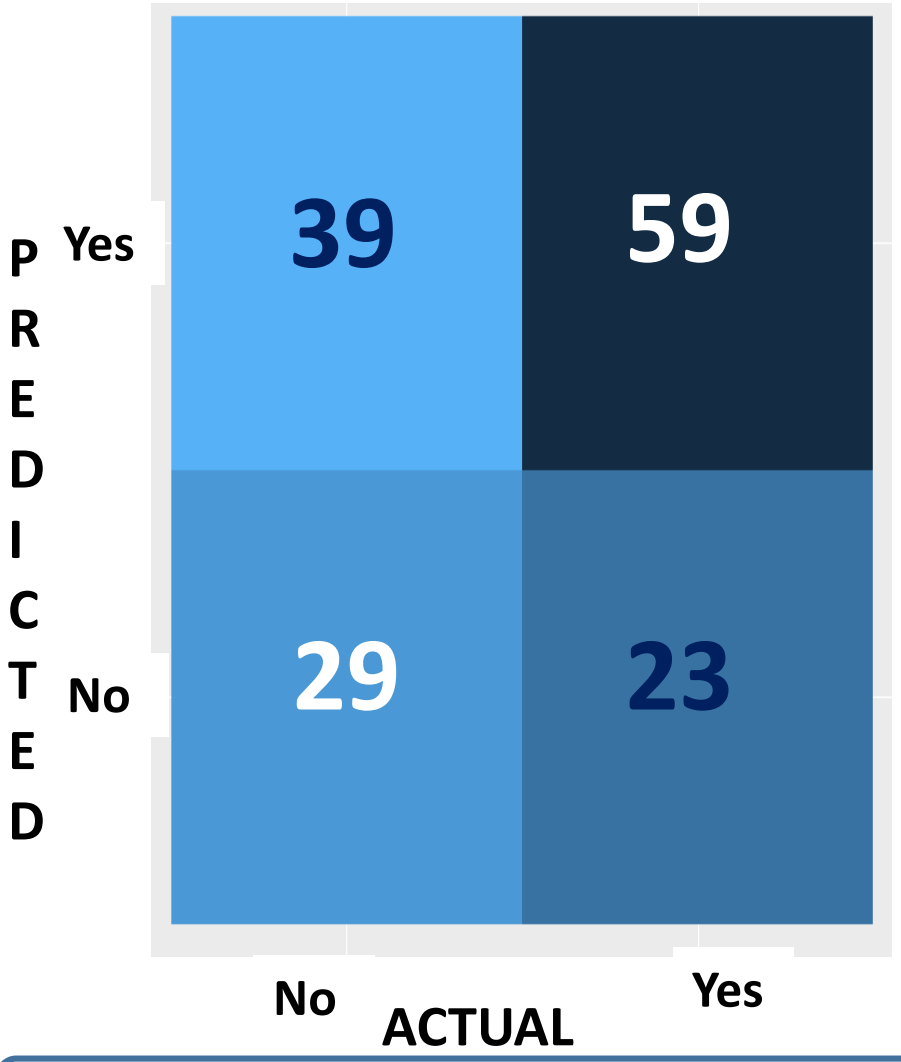
5-Fold Cross Validation to Determine the Optimal Number of Neighbors

The 5-fold CV was applied to the training data consisting of 603 High Schools



# of Neighbors

Overall Results of the KNN Model



Accuracy : 0.5867
95% CI : (0.5035, 0.6664)
No Information Rate : 0.5467
P-Value [Acc > NIR] : 0.18366
Sensitivity : 0.7195
Specificity : 0.4265
Pos Pred Value : 0.6020
Neg Pred Value : 0.5577
Prevalence : 0.5467
Detection Rate : 0.3933
Detection Prevalence : 0.6533
Balanced Accuracy : 0.5730

'Positive' Class : Yes

Method Overview: Classification and Regression Trees (CARTs)

- aka…*RECURSIVE Partitioning*
- For what problems can you apply this method?
  - Regression/Prediction for Continuous Outcome Variables
  - Classification for Categorical Outcome Variables (2 or more levels)
  - Tree methods can handle predictors that are continuous, categorical, skewed and sparse data as well as missing data
- Does this method have any Parameters?
  - Complexity Parameter (Cp)
  - Node Size/Branches/Fraction of Objects

How Do CARTs Work?

Step 1: Root Node Split. Starting with the entire data set (root node), select a variable (and corresponding range or grouping of levels) that generates two subsequent nodes that are each more "pure" than the root node.

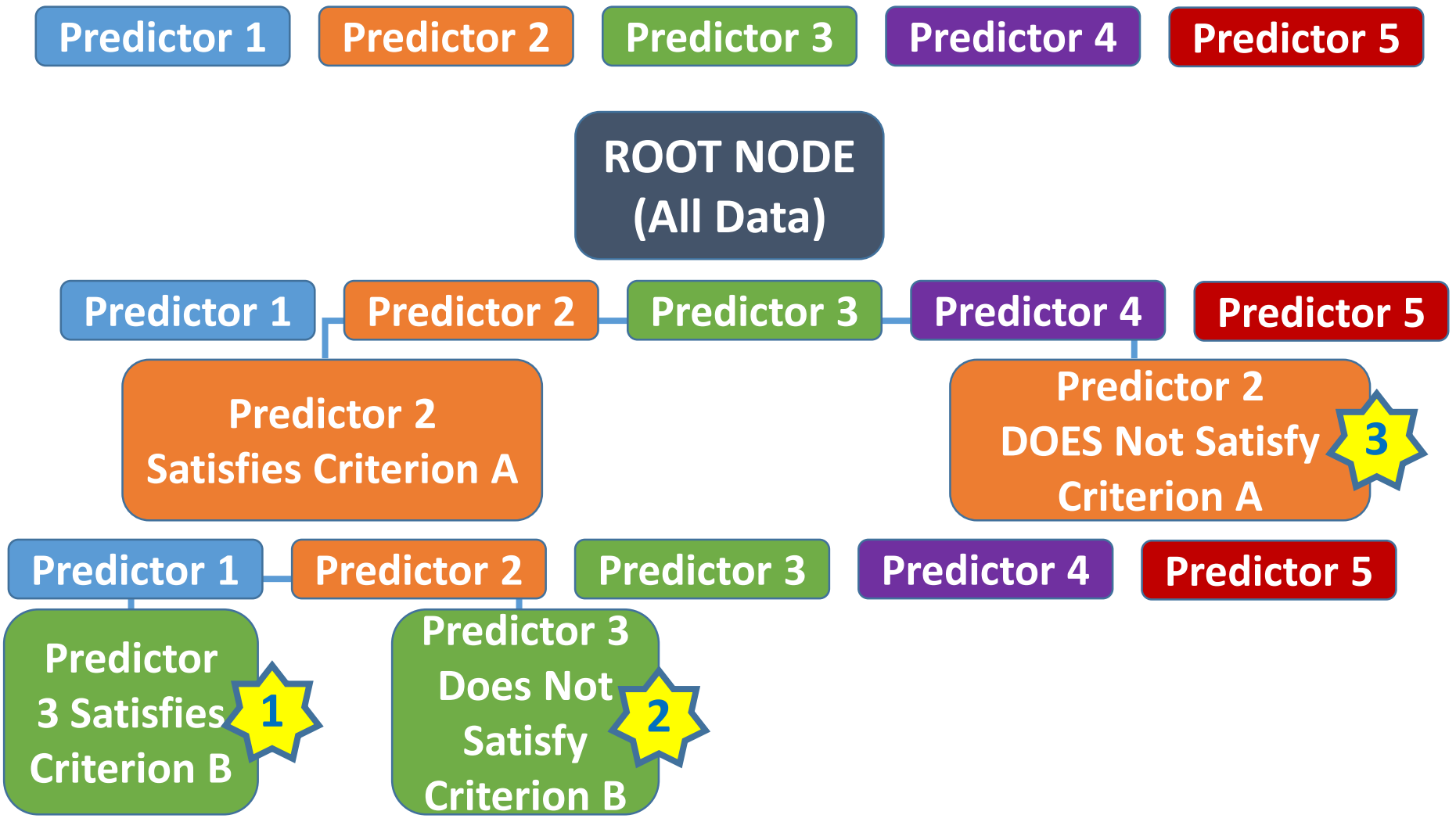
The usual optimization criterion for tree growing is Node Impurity is measured using the SSR (sum of squared residuals/errors) for continuous outcomes and using the Gini Index or Chi Squared/Deviance criterion for categorical outcomes.

Step 2: Continued Branching. Within each of the subsequent nodes, additional splits (branching) are determined from all possible predictors; any resulting child nodes are even more homogenous/pure than parent nodes.

Step 3: Pruning. After a tree is grown to completion, pruning branches back according to the complexity parameter is sometimes applied. Essentially, nodes and branches are removed from the tree producing a tree that is not as deep and with fewer final nodes.

Step 4: Prediction. Within each final node the predicted value is either the modal value/class of the outcome (Classification) or the mean of the outcome variable for observations in the node (Regression)

CART Overview



Method Overview: CARTs

- Insights on Parameters for this Method
  - Small Cp → lead to more complex trees (larger, deeper with more final nodes/leaves).
  - Large Cp → lead to less complex trees (smaller, more shallow with fewer final nodes/leaves)
- Caveats/ Considerations for the Method
  - Trees that are too large or overly complex can result in estimates that perform poorly for new data (low bias but high variance).
  - Conversely, trees that are not complex enough may result in estimates that don't adequately fit (lower variance but possibly larger bias).
  - Missing data are ignored when the tree is being built. However, when used for prediction, tree models can handle missing data by using SURROGATE variables – those related to the splitting variable and for which results are similar.
  - Predictors with large numbers of categories tend to get selected for splitting compared to those with fewer categories.

Method Overview: CARTs

- Advantages of this Method
  - Tree-based models are nonparametric and don't assume a functional form between outcome and predictors
  - Tree-based models naturally detect higher level interactions among the predictors.
  - Tree-based models are easy to visualize and interpret – even if the number of candidate predictors is large
- Disadvantages of this Method
  - Single trees are likely to have sub-optimal predictive performance compared to other methods (Kuhn and Johnson, 2013)
    - **This is especially true if the relationship between the outcome and predictors is not described well by "rectangular" boundaries**
  - The number of different predicted values (for regression trees) is determined by the number of final nodes in the tree.

Method Overview: CARTs

- Implementation of this method in R
  - rpart{rpart package}
  - tree{tree package};
  - ctree{party package}
  - Rattle {rattle package}
- Examples of this method in the S.R. Literature
  - Response Propensity Estimation and Nonresponse Bias Evaluation
    - **Phipps and Toth, 2012; http://1.usa.gov/1I4fObz**
    - **McCarthy et al. (2009); http://1.usa.gov/1IcQekK**
    - **Schouten and de Nooij (2005); http://bit.ly/1zUrPtf**
  - Frame Revision/Trimming
    - **McCarthy et al. (2009); http://1.usa.gov/1IcQekK**
  - Multiple Imputation
    - **Mesa, Tsai and Chambers (2000); http://bit.ly/1DrkSBt**
  - Generating Partially Synthetic Data
    - **Reiter (2005); http://bit.ly/1DY6kvz**

Trees and Forests for Survey Researchers

- http://bit.ly/BuskirkTreesandForests2018

Ensemble Methods

- Ensemble Methods refer to machine learning methods that combine multiple versions of the model together to make final predictions.
  - These methods generally have improved performance over single instances (less variance).
  - Sometimes the single model versions of the approach differ from the individual models grown in an ensemble to take advantage of the entire collection of models.
- Examples of Ensemble Methods Include:
  - random forests,
  - boosting and Bayesian additive regression trees (BART).
  - Extra Trees

Ensemble Methods

- One common way to create an ensemble method is to use the Bootstrap Aggregation Approach or Bagging, for short.
  - Bagging generates multiple versions of a model and uses these to get an aggregated estimate.
  - The aggregation averages over all the versions when predicting numerical outcome and relies on majority vote when predicting a categorical outcome.
  - Each version of the model is formed by taking a bootstrap sample the same size as the learning or training set.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2):123–140.

Method Overview: Random Forests

- Does this method have any Parameters?
  - Number of Trees (ntree)
  - # Variables to be randomly considered at each node (mtry)
  - Node sizes
- Insights on Parameters for this Method
  - Node size is typically 1 for classification and 5 (or more) for regression.
  - Generally, them more trees in the forest, the less error one has. But after a certain point, diminishing returns on error rate improvements are likely to be found in practice.
  - Larger values of mtry increases the strength of the tree, but also the correlation between the results of one tree from another in the forest.
  - Smaller values of mtry reduces the inter-tree correlation as well as the overall predictive strength of the tree.
    - **Default mtry is set as p/3 for regression**
    - **Default mtry is set as $\sqrt{p}$ (rounded down) for classification**

How Do Random Forests Work?

Step 1: Establish the Forest. Generate n_{tree} bootstrap samples from the original dataset with replacement and of the same size as the input dataset.

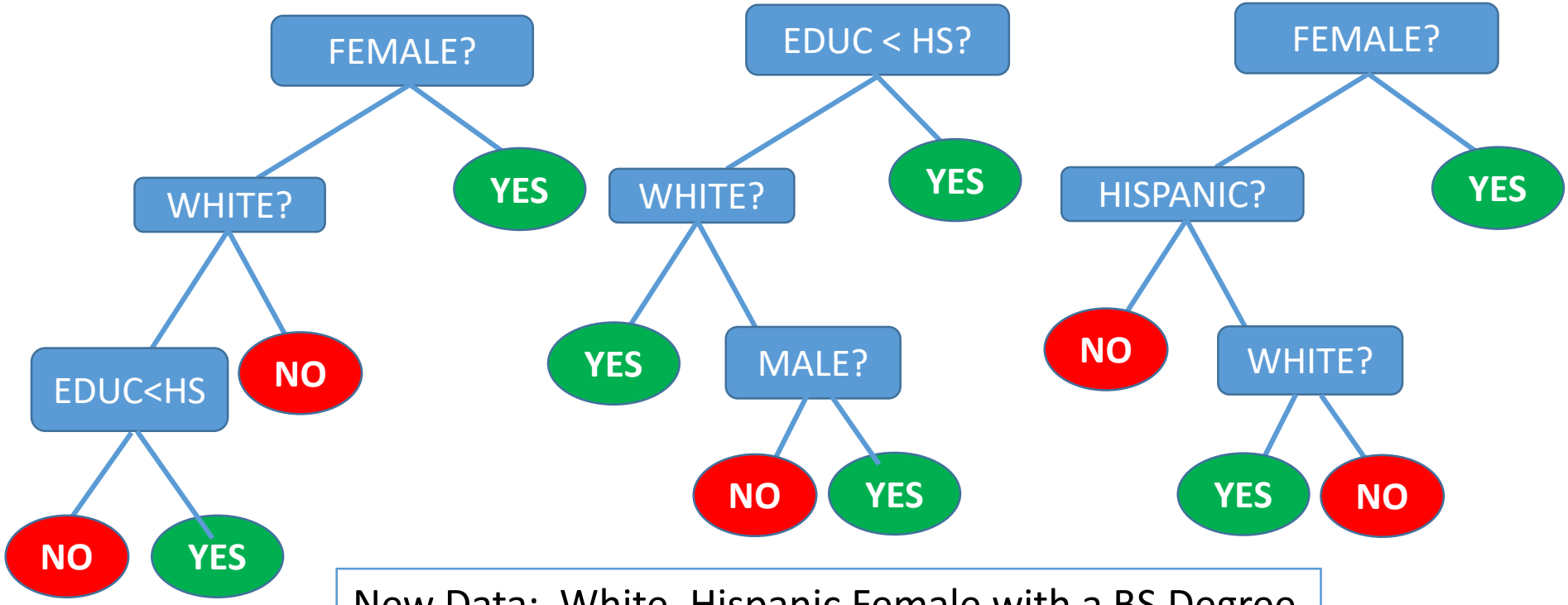
Step 2: Grow Trees. For each of the bootstrap samples drawn in Step 1, grow a classification/regression tree to full size (i.e. no pruning). But at **each** node of each of these trees, randomly select mtry of the predictor variables to be used as the basis of branching. Splits/branches are based on the best variable from among the mtry that are selected for a given node.

Step 3: Prediction. After all trees in the forest have been grown, form a prediction for new data from each tree in the same way as you would for normal classification or regression trees.

For forests based on regression trees, final predicted values for new data are computed as the mean of these n_{tree} predicted values;

For forests based on classification trees, final predicted values for new data are computed as the class with the majority vote across the ntrees (where each tree in the forest contributes one vote for a single outcome class).

Random Forests Illustrated…



New Data: White, Hispanic Female with a BS Degree

Tree 1: YES

Tree 2: YES

Tree 3: NO

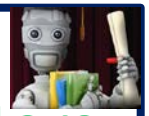
Forest Uses Majority Rule: Prediction for New Data: YES

Method Overview: Random Forests

Caveats/ Considerations for the Method

- For situations where the number of predictors is greater than the number of cases (n<<p), mtry values should be increased beyond the defaults.
  - **See: https://samos.univ-paris1.fr/IMG/pdf_S1E1-Poggi.pdf**
- Random Forests do not produce coefficients or significance tests for predictors. They do produce variable importance measures that can help rank order the utility of the variables in predicting the outcome.
- Random Forests don't use surrogates like trees, so missing data must be dealt with
- Variable Importance measures can be biased if the predictor variables are either highly correlated or on different scales.
  - **Suggest using Conditional Forests instead (Strobl et al., 2007)**
  - **See: http://www.biomedcentral.com/1471-2105/8/25**

Method Overview: Random Forests

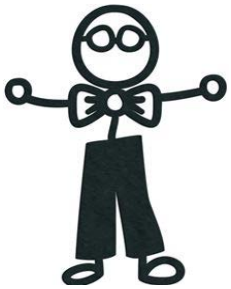


- Advantages of this Method
  - Like Tree methods, Random Forests can handle predictors that are continuous, categorical, skewed and sparse data as well as missing data.
  - Random forests are also aptly suited for the "large p, small n" scenario (Strobl et al., 2007).
  - Random forests can also be very effective for estimating outcomes that are a complex functions of predictors with many interactions or possibly non-linear functions of the parameters (Mendez et al., 2008).
- Disadvantages of this Method
  - Random Forests can be computationally intensive.
  - Unlike Trees, Random Forests are not easily visualized.

Method Overview: Random Forests

- Implementation of this method in R
  - randomforest{randomForest package}
  - rfsrc{randomForestSRC + ggRandomForest packages}
  - cforest{party package}
  - CoreModel{CORElearn package}
- Examples of this method in the S.R. Literature
  - Response Propensity Weighting and Nonresponse Bias Evaluation
    - **Buskirk and Kolenikov (2015); http://surveyinsights.org/?p=5108**
    - **Buskirk, West and Burks (2013); http://bit.ly/1JVzgWt**
  - Hybird Ensemble Method for Response Propensity Scores and Nonresponse Bias evaluation for Establishment Surveys
    - **Earp, Mitchell, McCarthy and Kreuter (2014); http://bit.ly/1Pi4GIf**
  - Identifying important predictors of county-level CPO rates
    - **Buskirk, et al. (2014); http://bit.ly/1Jd3TXt**

My First Journey through the Forest...



+

=

01000001010000000 Learning about

Factors Associated With Persistence in Science and Engineering Majors: An Exploratory Study Using Classification Trees and Random Forests

GUILLERMO MENDEZ
Department of Mathematics and Statistics
Arizona State University

TRENT D. BUSKIRK
School of Public Health
Saint Louis University

SHARON LOHR
Department of Mathematics and Statistics
Arizona State University

SUSAN HAAG
Ira A. Fulton School of Engineering
Arizona State University

ABSTRACT

Many students who start college intending to major in science or engineering do not graduate, or decide to switch to a non-science major. We used the recently developed statistical method of random forests to obtain a new perspective of variables that are associated with persistence to a science or engineering degree. We describe classification trees and random forests and contrast the results from these methods with results from the more commonly used method of logistic regression. Among the variables available in Arizona State University data, high school and freshman year GPAs have highest importance for predicting persistence; other variables such as number of science and engineering courses taken freshman year are important for subgroups of the student population. The method used in this study could be employed in other settings to identify faculty practices, teaching methods, and other factors that are associated with high persistence to a degree.

Keywords: classification tree, logistic regression, random forest

I. INTRODUCTION

Many studies have shown a lack of persistence among U.S. students who complete a science and engineering degree (Besterfield-Sacre, Atman, and Shuman, 1997; Brainard and Carlin, 1997; Burtner, 2005; Grandy, 1998; May and Chubin, 2003; LeBold and Ward, 1998; Leslie, McClure, and Oaxaca, 1998; Levin and Wyckoff, 1991; Rayman and Brett, 1995; Seymour and Hewitt, 1997; White, 2005; Zhang, Anderson, Ohland, and Thorndyke, 2004). These studies have identified a number of variables such as high school GPA that are associated with persistence to a degree. Most previous work has identified factors related to persistence using standard statistical methods such as logistic regression. These methods work well for identifying simple relationships in the data. However, when predicting whether a student will graduate with an engineering degree, the relationships are often more complex. For example, female Hispanic students who participate in a mentorship program are more likely to persist to a degree, while some other groups of students in the program are less likely to persist. Such a relationship is easily missed when techniques such as logistic regression are used.

In this paper we use classification trees (Breiman, Friedman, Olshen, and Stone, 1984) to produce a new view of variables associated with persistence to earn a science, technology, engineering, or mathematics (STEM) degree. We also use the recently developed statistical method of random forest (Breiman, 2001), related to tree-based classification methods, to identify factors that may be related to persistence but that might not be identified by other statistical procedures such as logistic regression. The primary goal of this paper is to show how classification trees and random forests can be used to identify factors and interactions not found by other methods.

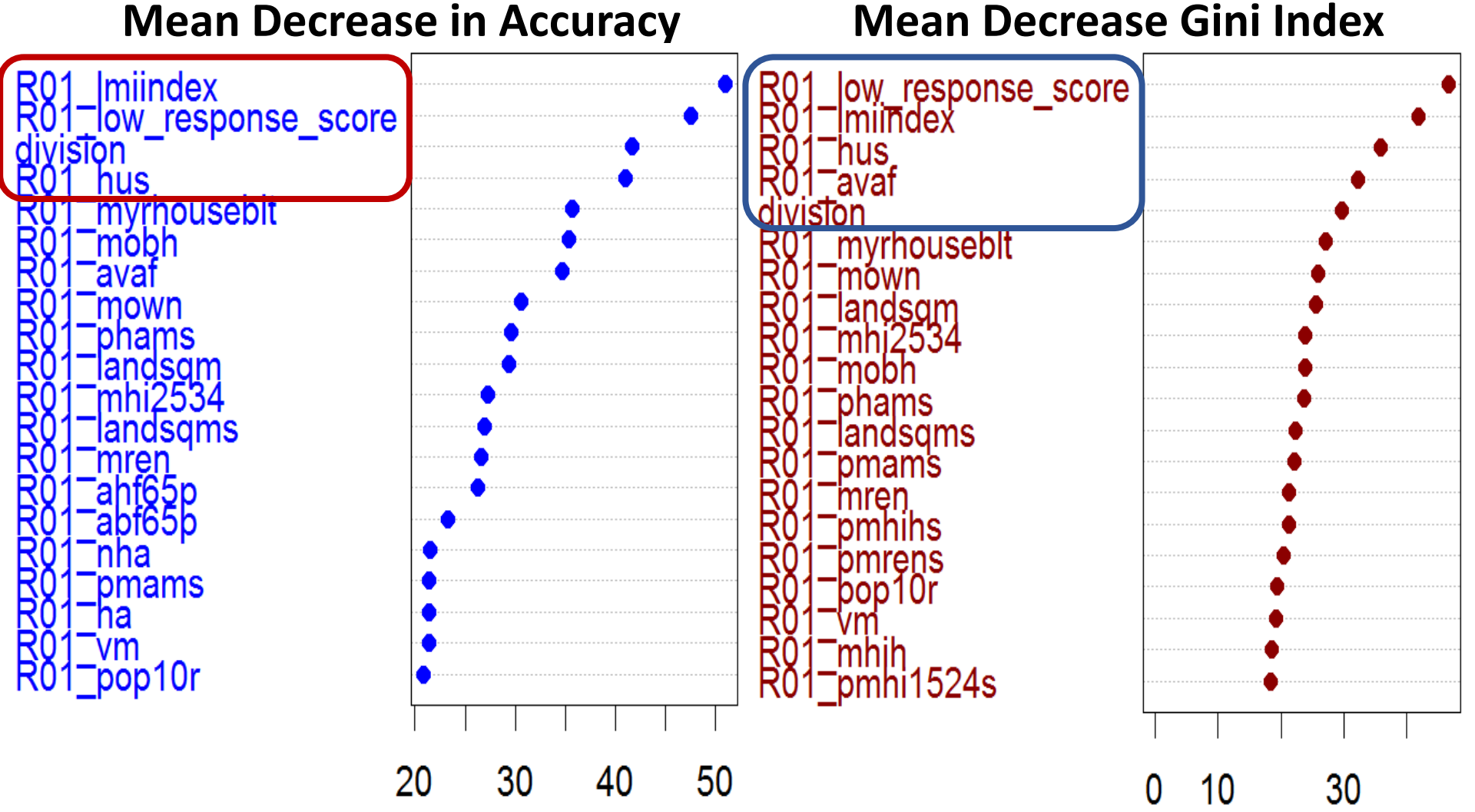
Zhang et al. (2004) suggested that high school GPA and SAT math scores predicted engineering student graduation. However, these two cognitive variables explained only a small fraction of the overall variability in student graduation persistence rates suggesting that more predictors are needed to fully understand the nature of persistence in science and engineering. A recent study by Burtner (2005) supports the use of non-cognitive variables, such as confidence in college-level math/science ability, in models to predict student persistence. Other studies (Besterfield-Sacre, Atman, and Shuman, 1997; Brainard and Carlin, 1997) have supported Burtner's assertions by demonstrating associations between graduation rates and attitudinal and belief factors such as self-confidence and perceived ability in engineering as well as other factors such as work status, high school ranking, and SAT scores. Levin and Wyckoff (1991) also reported that high school GPA, scores on college placement tests in Chemistry, along with grades in Calculus, Chemistry, and Physics courses were all strong predictors of persistence through the second year of engineering programs. LeBold and Ward (LeBold and Ward, 1988) found that first and second semester grades along with cumulative GPA were strong predictors of persistence for freshmen engineering majors.

The majority of studies investigating persistence in science and engineering have focused on engineering students. Enrollment and tracking of engineering majors may be two key factors related to the

Persisting through the Forest

- In this study we looked at a cohort of University Graduates who declared a STEM major during their freshmen year. We wanted to understand the influential factors related to predicting STEM persistence…
- The models were based on Random Forests (a relatively new machine learning method for classification and prediction) based on data that included:
  - University Administrative Records
  - Student Background/Demographic Data
  - Student GPA data (H.S. and College)
  - Student Major and Course related information
  - High School GPA and SAT test scores
  - Other enrollment related variables during freshmen year

Variable Importance for Random Forest Example



Machine Learning Algorithms

Select machine learning techniques

- KNN
- Ensemble Methods
- Random Forests
- **Extra Trees**

Method Overview: Extremely Randomized Trees (or Extra-Trees, for short)

- Does this method have any Parameters?
  - Number of Trees (M)
  - # Variables to be randomly considered at each node (K)
  - Min Node sizes (nmin)
  - Number of Cut Points (continuous) and NumBins (Categorical)
- Insights on Parameters for this Method
  - Smaller values of K increases the level of randomness in the process. Geurt et al. (2006) observed that bias monotonically decreased, but variance increased with increases in K.
    - **Default K is set as p/3 (rounded down) for regression**
    - **Default K is set as $\sqrt{p}$ (rounded down) for classification**
  - The number of cut points for continuous variables is 1, by default, but can be increased. As NumBins increases so do the risks for overfitting.
  - Node size is typically 2 for classification and 5 (or more) for regression. Larger values of this hyperparameter lead to smaller, more biased trees but smaller variance.

How Do Extra-Trees Work?

Step 1: Create a Tree. Generate a regression or classification tree using the entire (training) data set as in the CART algorithm, but at each node only select K variables to be considered for node splitting (as in Random Forests). At each node randomly select numcutpoints for each continuous predictor and numbins for each categorical variable. From among these implied partitions, select the best variable and split point for creating more homogeneous nodes.

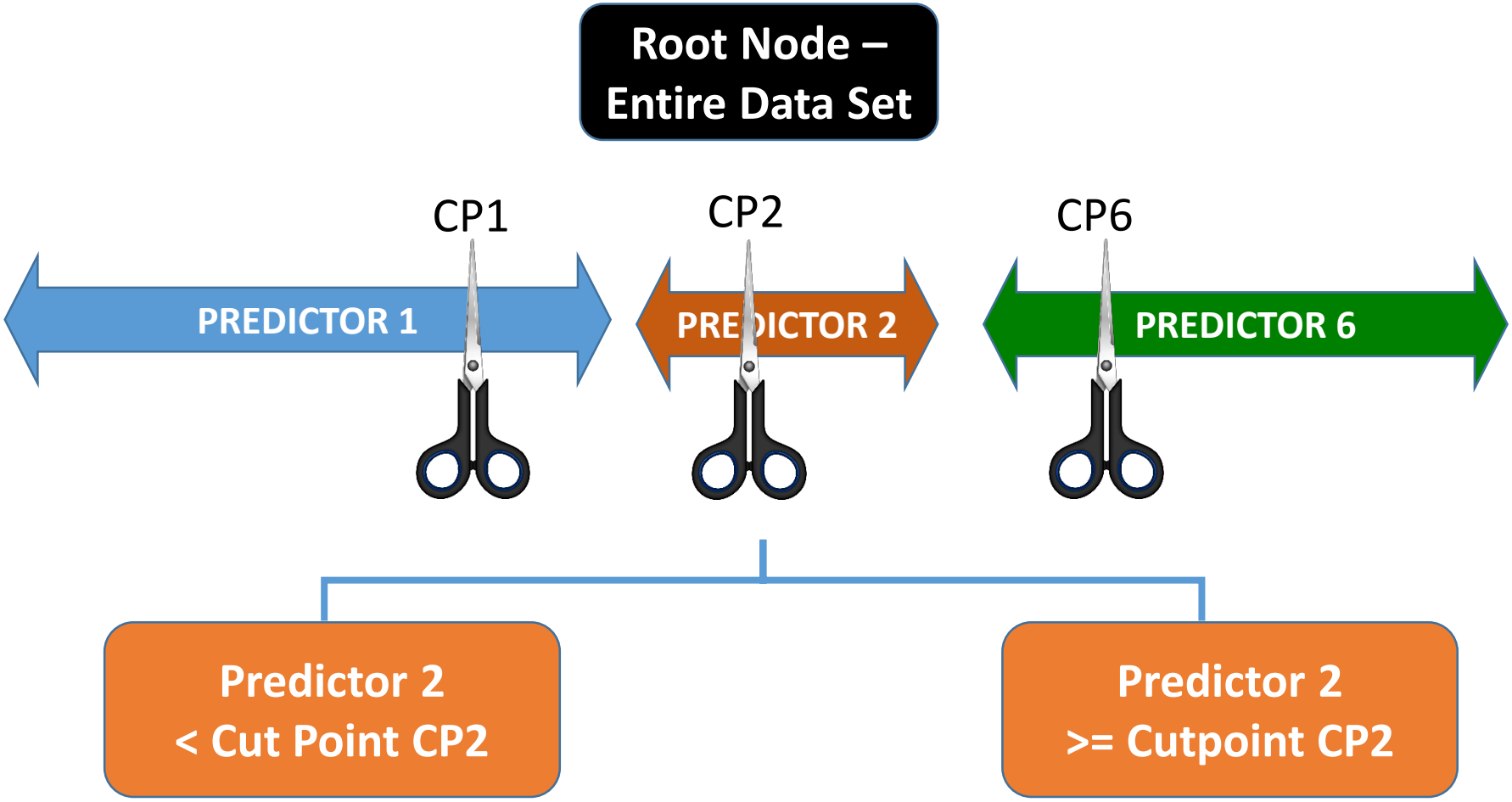
Step 2: Stop Growing the Tree. Continue the (binary splitting) process until the final node sizes cannot be split further because they contain fewer cases than the minimum node size (nmin). There is no pruning.

Step 3: Create Extra Trees. Repeat steps 1 and 2 a total of M-1 additional times to create an ensemble of M extra trees.

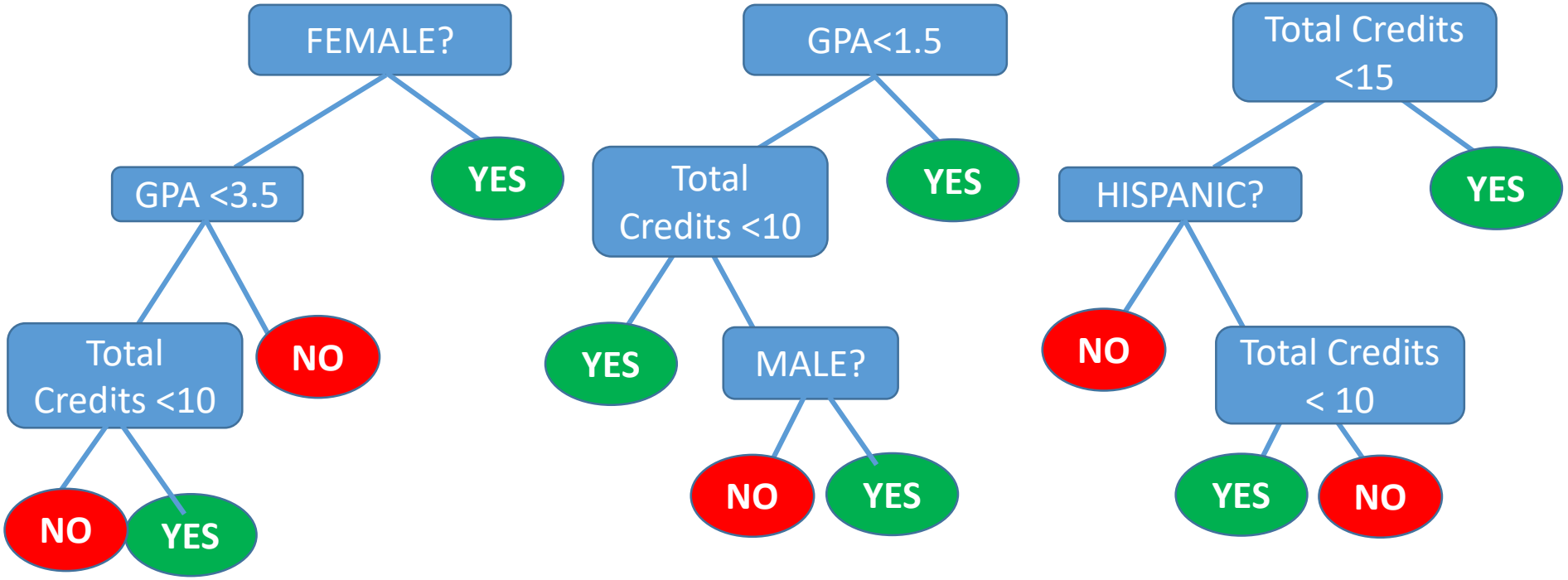
Step 4: Create Predictions. For extra regression trees, final predicted values for new data are computed as the mean of the M predicted values;

For Extra-Trees based on classification trees, final predicted values for new data are computed as the class with the majority vote across the M trees (where each tree in the ensemble contributes one vote for a single outcome class).

Node Splitting for Extra Trees (based on continuous predictor)



Extra Trees Illustrated…

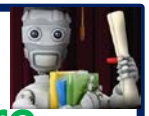


New Data: Hispanic Female taking 12 credits with a current GPA of 3.0

Tree 1: YES **Tree 2: YES** **Tree 3: NO**

Extra-Trees Uses Majority Rule: Prediction for New Data: YES

Method Overview: Extra Trees



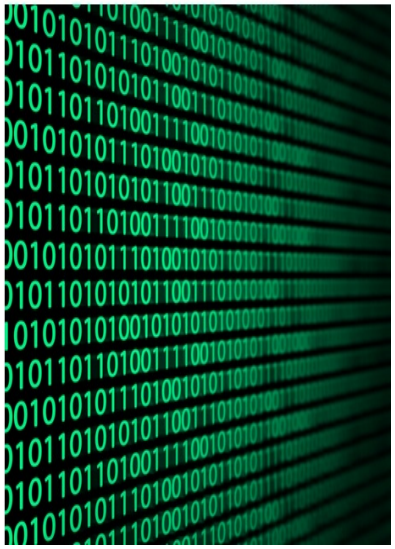
- Advantages of this Method
  - Like Tree methods, Extra-Trees can handle predictors that are continuous, categorical, skewed and sparse.
  - Extra-Trees are computationally faster than other ensemble methods and may be more aptly suited for very large data sets (either in terms of number of cases or number of variables).
  - Extra-trees generally reduce the variance in estimates and often the this reduction in variance exceeds increases in bias (Geurts et al. 2006).
- Disadvantages of this Method
  - Extra-Trees are sensitive to the number of bins for categorical predictors and higher values may lead to increases in bias and computational time.
  - Extra-Trees can create more complex trees compared to other tree-based methods and this trait has been linked to slight increases in bias.

Examples of Extra-Trees

- Kern et al. (2018) compare tree based methods, including random forests and extra trees, among others, for predicting nonresponse within a survey panel context.

- Kern, Weiß, Kolb (2019) extend this investigation to compare prediction of panel attrition based on substantive survey data as well as panel participation data.

- Extra-trees can be implemented in R through:
  - extraTrees package (modelled after the randomForest package)
  - h20 pacakge (see https://www.h2o.ai/ for more info)



Resources for Machine Learning

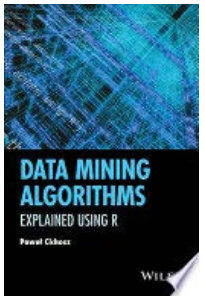
Training

R meta packages for ML

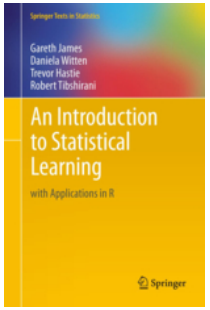
Other Open Source and Proprietary Resources

Data Mining/Machine Learning Resources

- http://www.dataminingconsultant.com/resources.htm
- Data Mining Algorithms Explained Using R (2015)
  - http://bit.ly/1yZYHjK



- Data Mining for the Social Sciences (2015)
  - http://bit.ly/1DpPFC2
- An Introduction to Statistical Learning with Applications in R (2013)
  - Free PDF Version: http://bit.ly/1iUJso0
  - Online Resources for FREE lecture videos and labs in R
    - **http://bit.ly/1snBMk5**

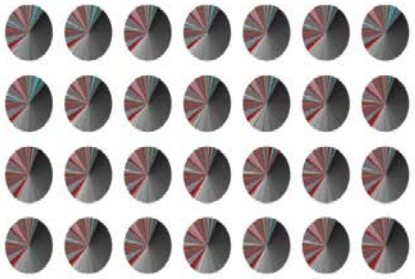


- An overview of Machine Learning Functions available in R
  - http://cran.r-project.org/web/views/MachineLearning.html

Resources for Learning More…

- Please Check out the Volume 11, Issue 1: Jan, 2018 of Survey Practice dedicated to Introducing Machine Learning Methods to Social and Survey Researchers http://bit.ly/SPMachineLearningIssue

ARTICLES



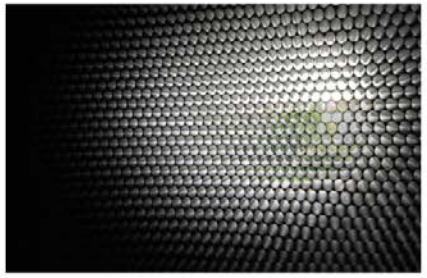
Using Support Vector Machines for Survey Research

Antje Kirchner, Curtis S. Signorino

Recent developments in machine learning allow for flexible functional form estimation beyond the approaches typically used by survey researchers and social scientists. Support vector machines (SVMs) are one such …

Abstract

ARTICLES



Using LASSO to Model Interactions and Nonlinearities in Survey Data

Curtis S. Signorino, Antje Kirchner

The LASSO and its variants have become a core part of the machine learning toolkit. Similar to OLS and logistic regression, the LASSO can be applied to continuous …

Abstract

ARTICLES

Neural Networks for Survey Researchers

Adam Eck

Neural networks are currently one of the most popular and fastest growing approaches to machine learning, driving advances in deep learning for difficult real-world applications ranging from image recognition …

Abstract

ARTICLES

Surveying the Forests and Sampling the Trees: An overview of Classification and Regression Trees and Random Forests with applications in Survey Research

Trent D. Buskirk

While survey and social science researchers have become well versed in traditional modeling approaches such as multiple regression or logistic regression, there are more contemporary …

Abstract

Meta Packages for Machine Learning in R

- There are several "meta" packages in R for machine learning that bundle many algorithms together into a unified interface, ecosystem or dashboard that enables researchers and analysts to create several machine learning models efficiently.
- Several meta packages are avaialble in R including:
  - Caret – one of the first such packages
  - mlr - offers a unified common language for processing data, developing ML models and evaluating and visualizing them. Think grammar of graphics for ML.
    - **https://mlr.mlr-org.com/**

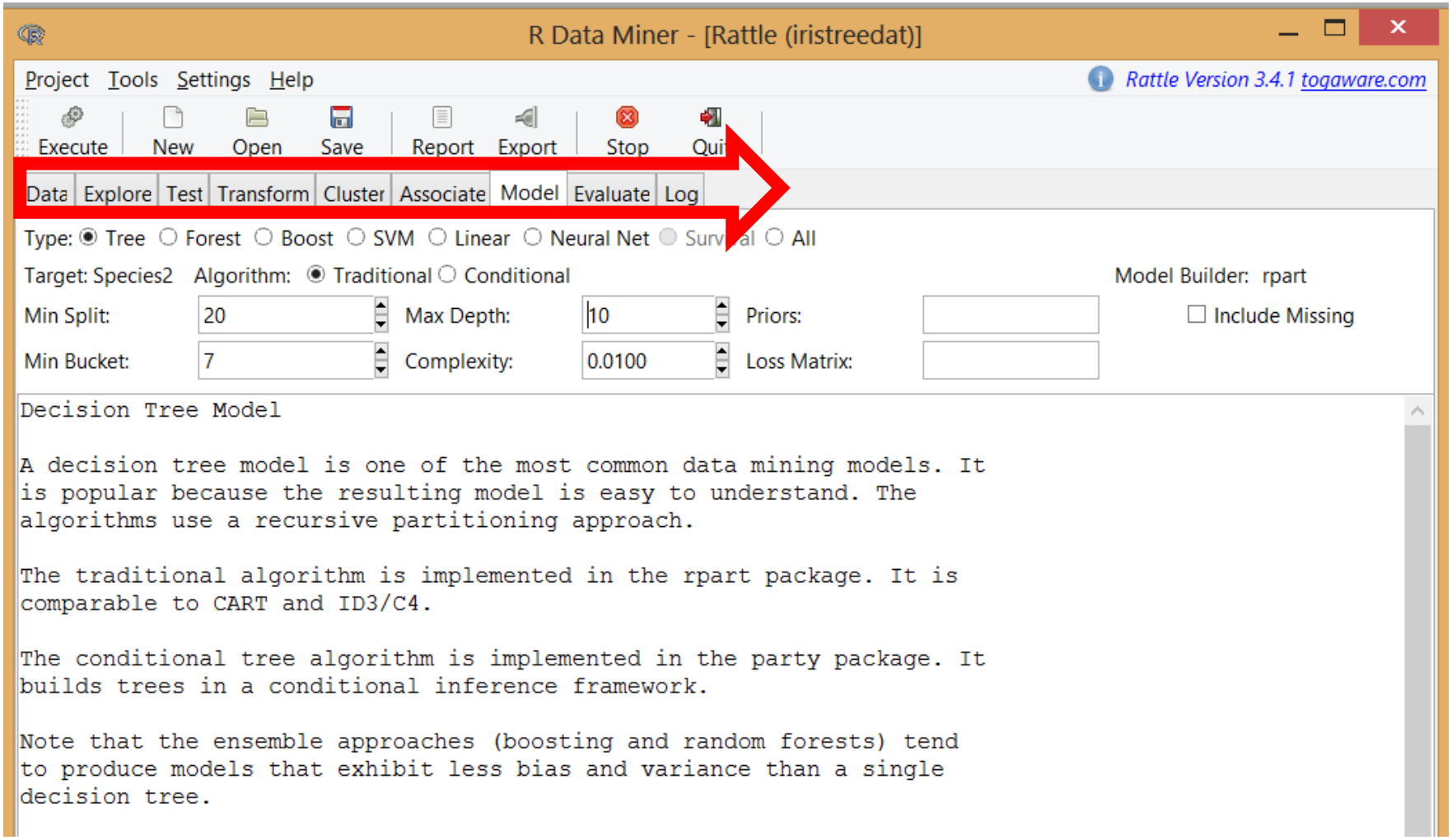
Meta Packages for Machine Learning in R

- Several meta packages are avaialble in R including:
  - Rattle – the rattle package in R provides graphical user "point-and-click" interface for many machine learning methods and creates a log file that generates syntax required for models requested through the GUI interface.
- There are also free-standing open-source and paid software meta-machine learning packages that help automate the ML model development process:
  - DataRobot – just topped Forbes AI list as one of the most well-funded AI Automation startups.
    - **https://www.datarobot.com/**
  - H20 – similar freestanding software to DataRobot but has R packages and python libraries that can be used to call the software capabilities from within R/python environment.
    - **https://www.h2o.ai/**

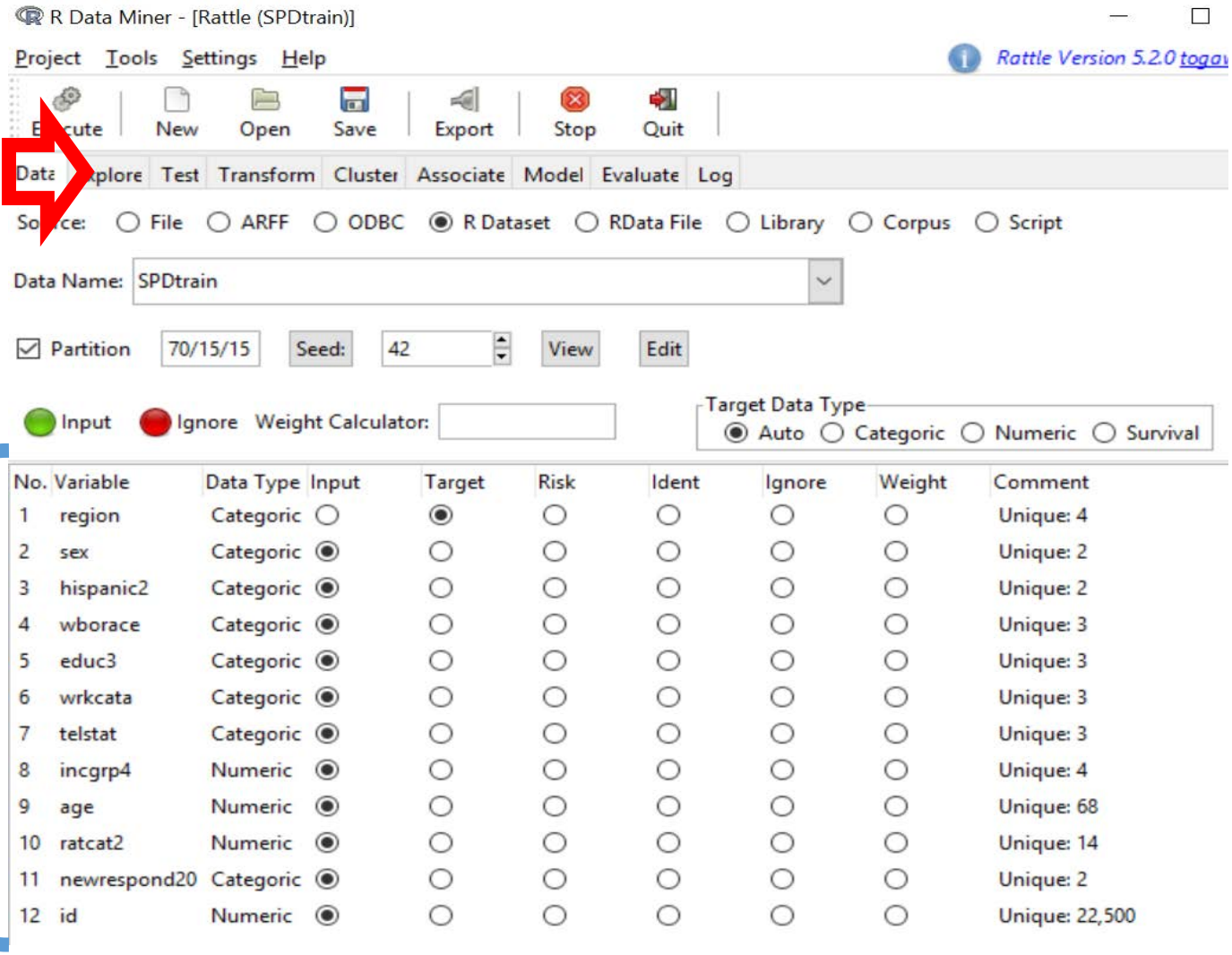
A Demonstration of the Rattle Package


```
install.packages("rattle", dependencies=TRUE)
library(rattle)
rattle()
```

Machine Learning in R via Rattle!

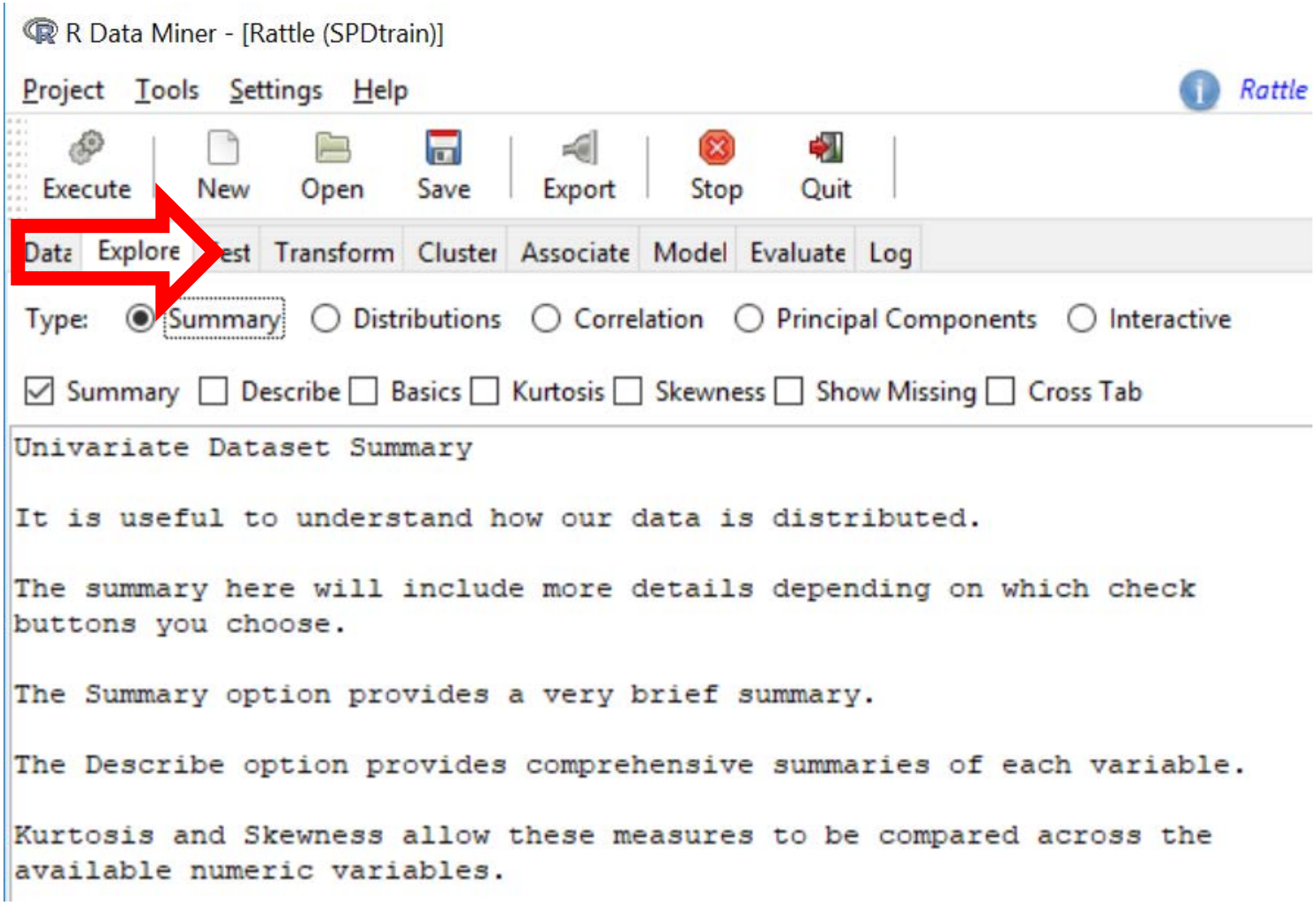


Data Tab in Rattle

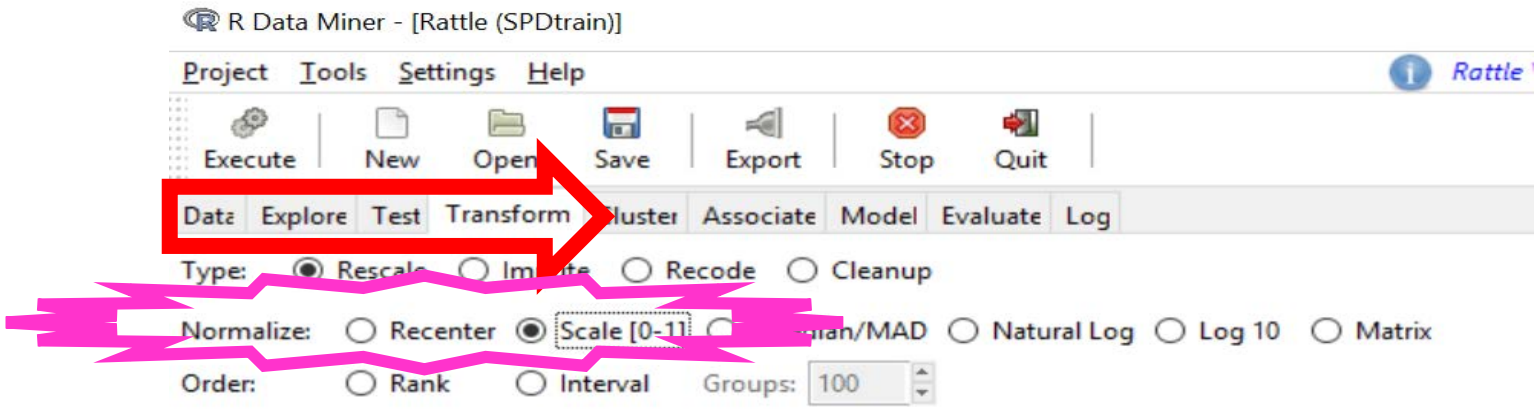


No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	region	Categoric	○	●	○	○	○	○	Unique: 4
2	sex	Categoric	●	○	○	○	○	○	Unique: 2
3	hispanic2	Categoric	●	○	○	○	○	○	Unique: 2
4	wborace	Categoric	●	○	○	○	○	○	Unique: 3
5	educ3	Categoric	●	○	○	○	○	○	Unique: 3
6	wrkcata	Categoric	●	○	○	○	○	○	Unique: 3
7	telstat	Categoric	●	○	○	○	○	○	Unique: 3
8	incgrp4	Numeric	●	○	○	○	○	○	Unique: 4
9	age	Numeric	●	○	○	○	○	○	Unique: 68
10	ratcat2	Numeric	●	○	○	○	○	○	Unique: 14
11	newrespond20	Categoric	●	○	○	○	○	○	Unique: 2
12	id	Numeric	●	○	○	○	○	○	Unique: 22,500

Explore Tab in Rattle

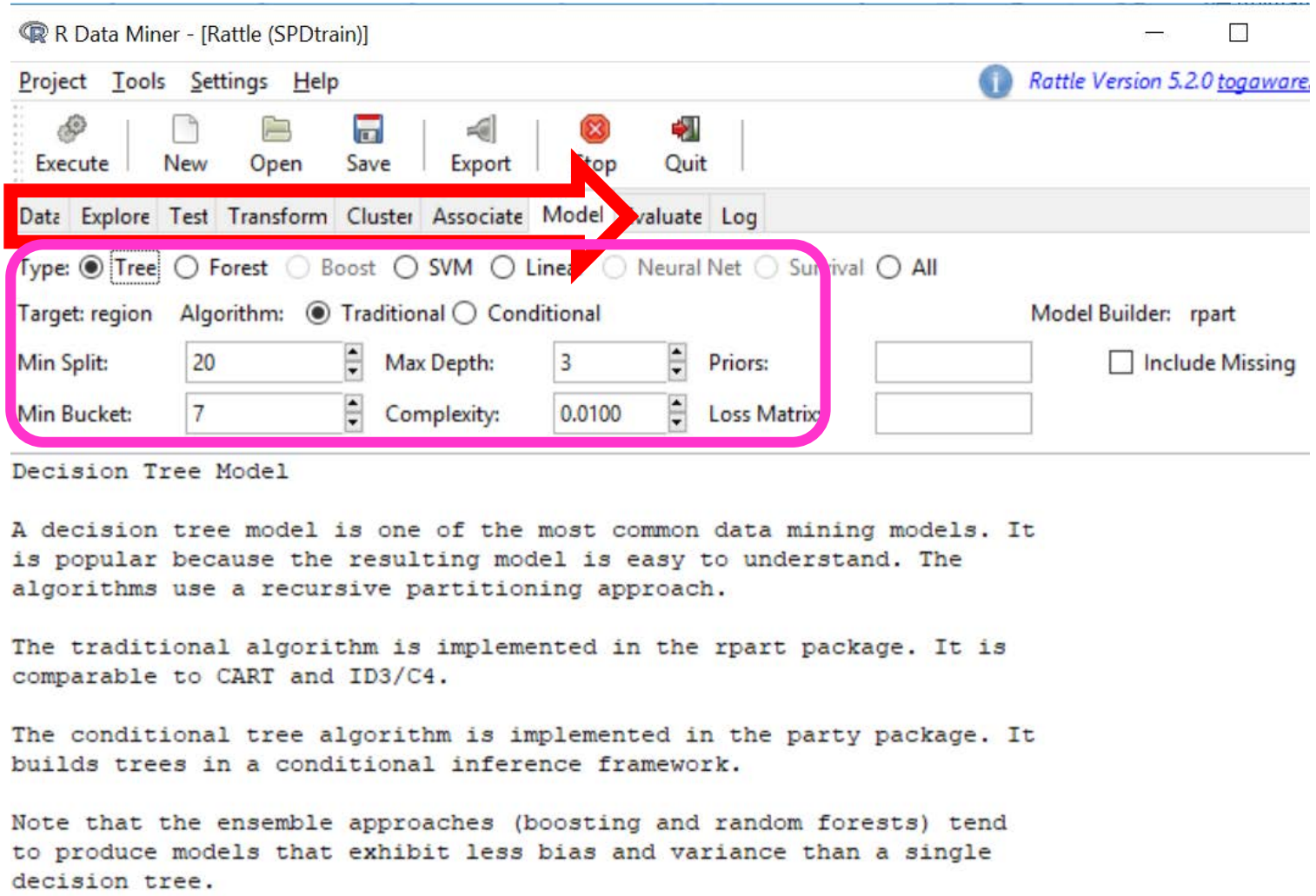


Transform Tab in Rattle

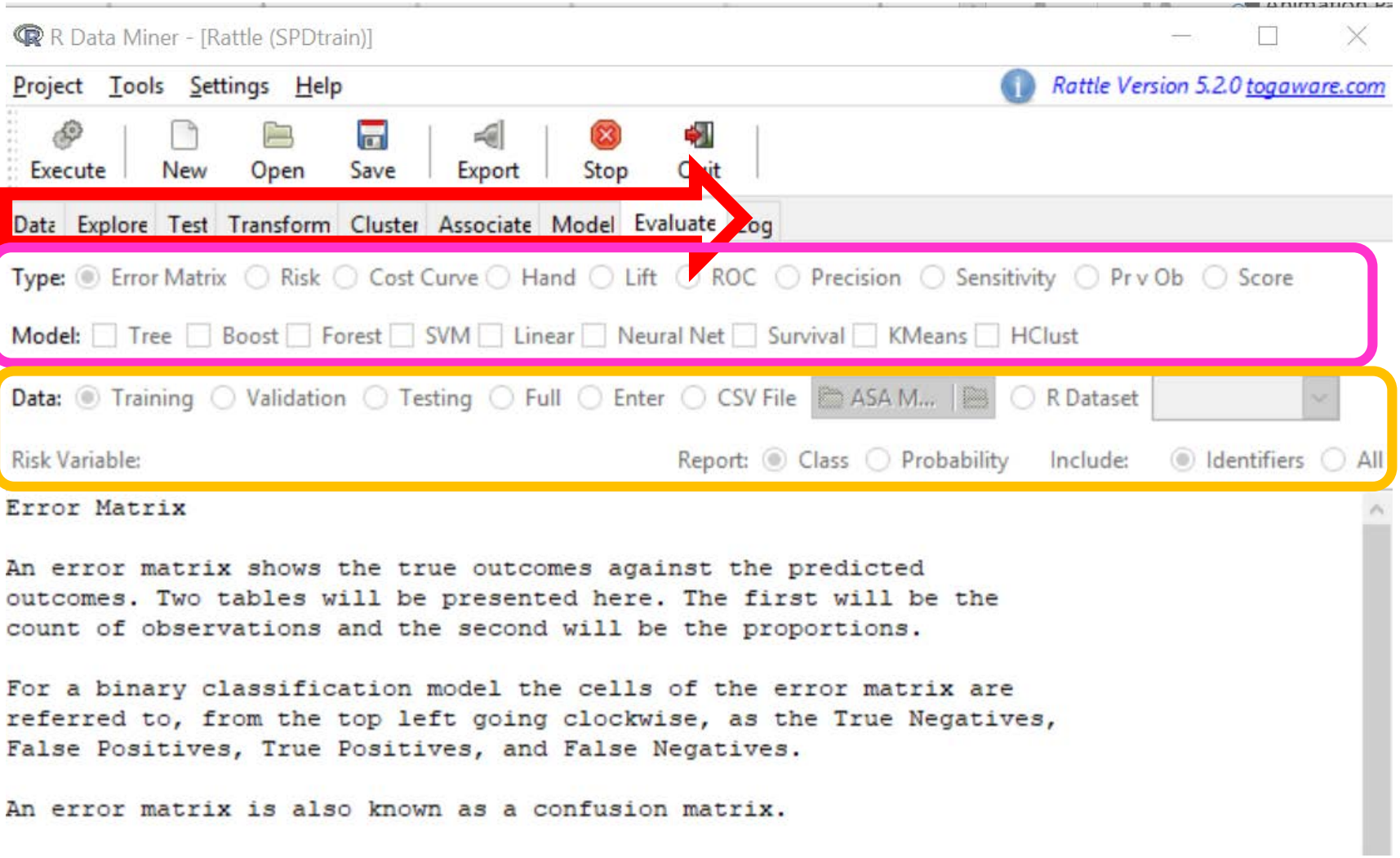


No.	Variable	Data Type and Number Missing
1	region	Categorical [4 levels].
2	sex	Categorical [2 levels].
3	hispanic2	Categorical [2 levels].
4	wborace	Categorical [3 levels].
5	educ3	Categorical [3 levels].
6	wrkcata	Categorical [10 levels].
7	telstat	Categorical [3 levels].
8	incgrp4	Numeric [1.00 to 4.00; unique=4; mean=2.18; median=2.00].
9	age	Numeric [18.00 to 85.00; unique=68; mean=47.74; median=47.00].
10	ratcat2	Numeric [1.00 to 14.00; unique=14; mean=8.54; median=9.00].
11	newrespond20	Categorical [2 levels].
12	id	Numeric [1.00 to 26785.00; unique=22500; mean=13378.68; median=13413.50].

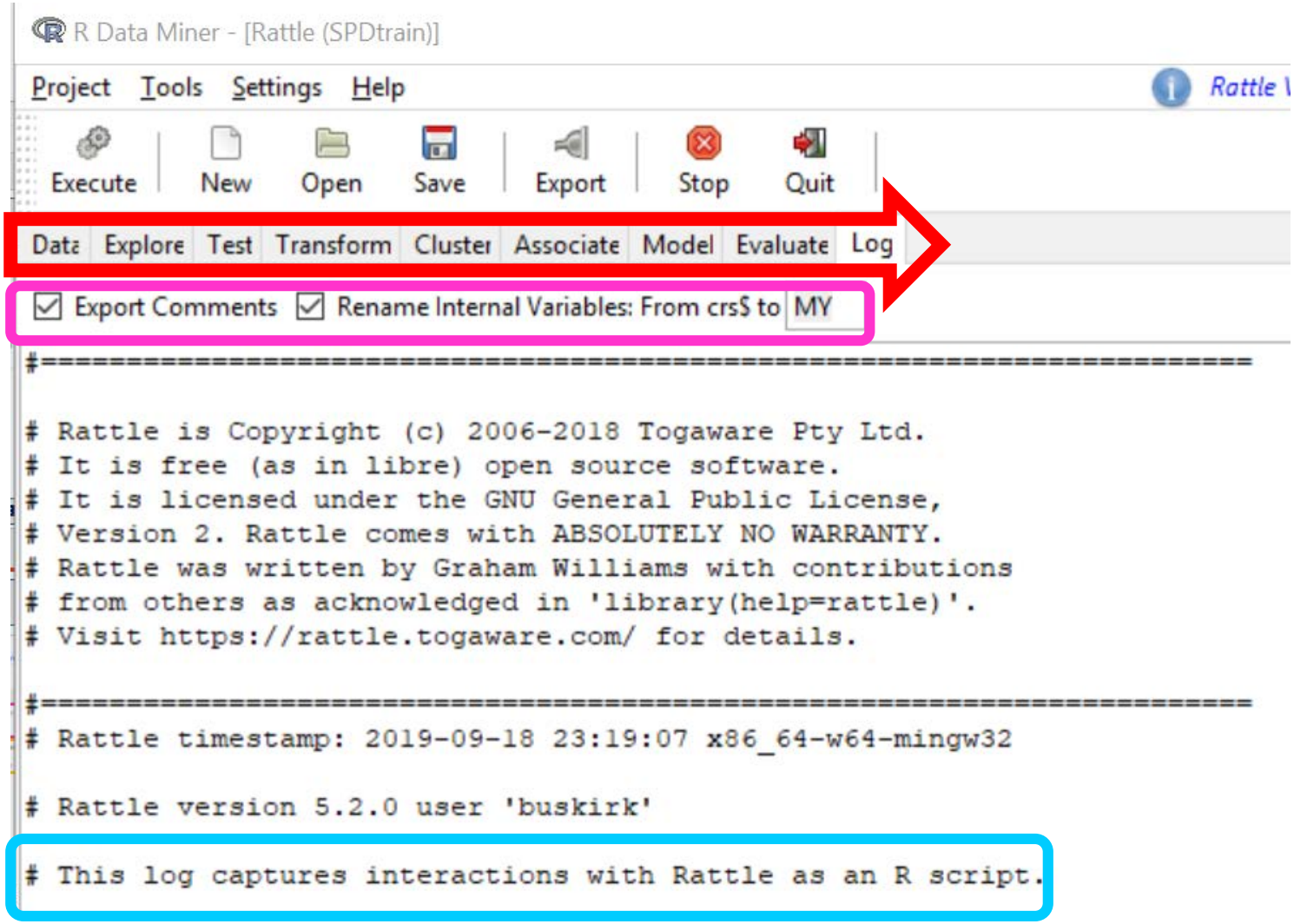
Model Tab in Rattle



Evaluate Tab in Rattle



Log Tab in Rattle



CART EXAMPLE

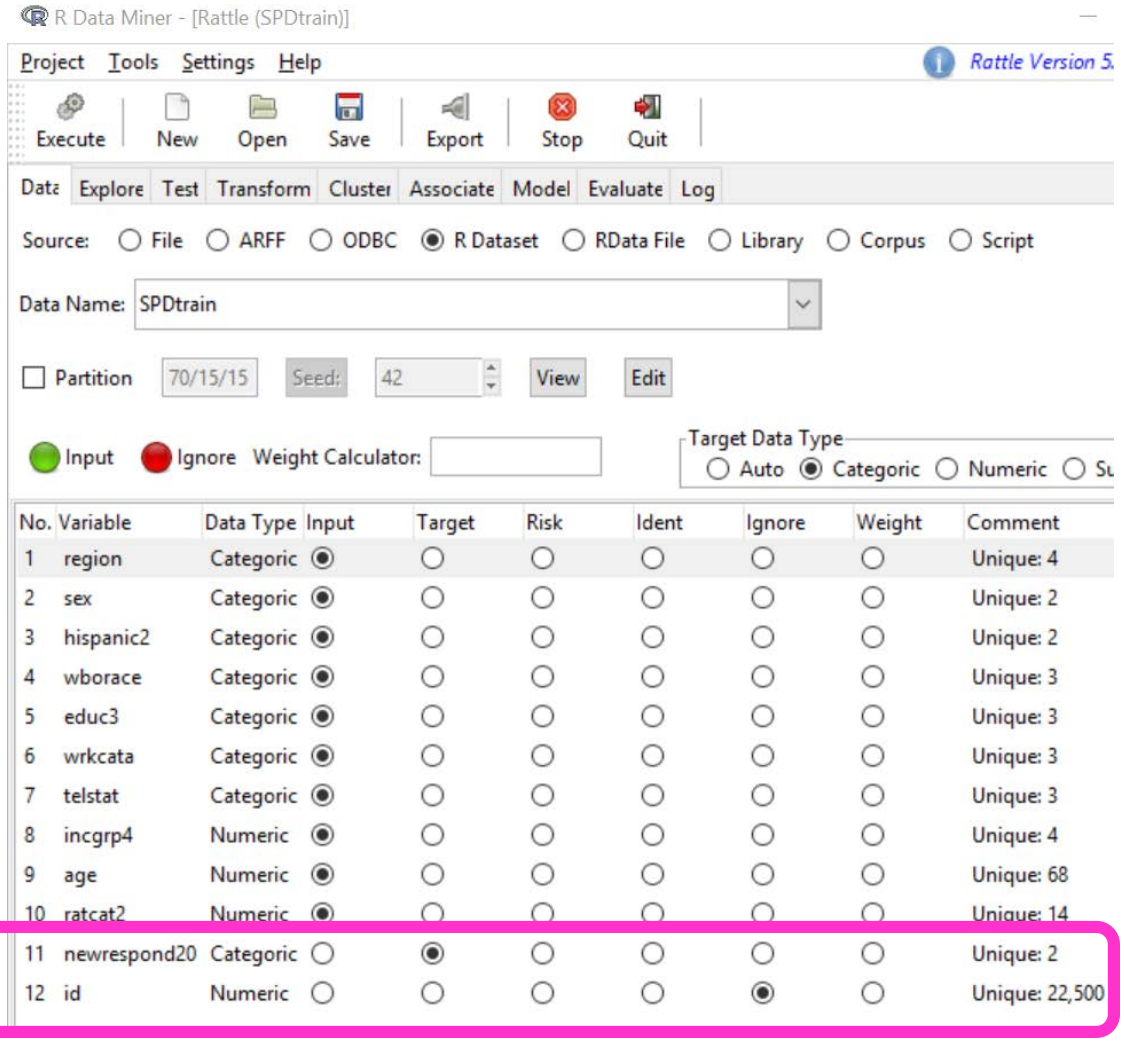
- Example CART model exercises and typical solution is available in the webinar handout, data file and script file.

- See:
  - BuskirkASAWebinarExample.pdf
  - ASA2019LabData.RData
  - ASA2019LabData.R

CART Example in Rattle

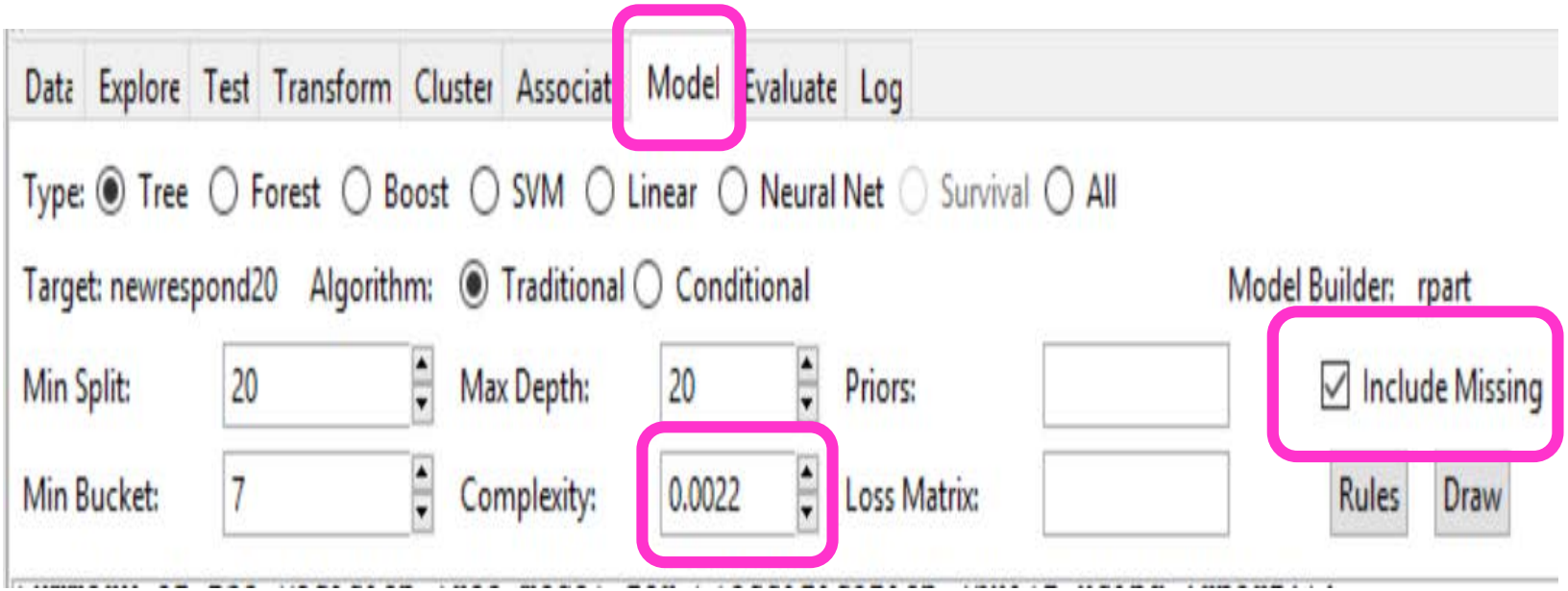
- The ASA2019TreeData.RData also contains **the data SPDtrain and SPDtest which we will use for this part to build a tree model to predict response status from a collection of covariate values that have been described in this article:** https://www.surveypractice.org/article/2718-an-introduction-to-machine-learning-methods-for-survey-researchers.
- **The outcome of interest is newrespond20 and the predictors we want to use are all of the demographic variables including age, income, race, etc. that are listed in the table in the article referenced above. The id number is not a predictor, but is included in the file just for reference.**

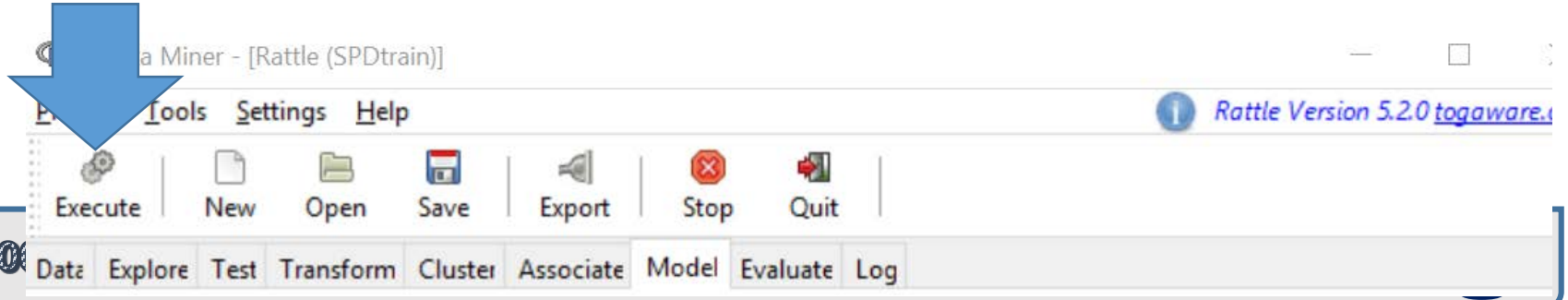
Step 1: Process Data



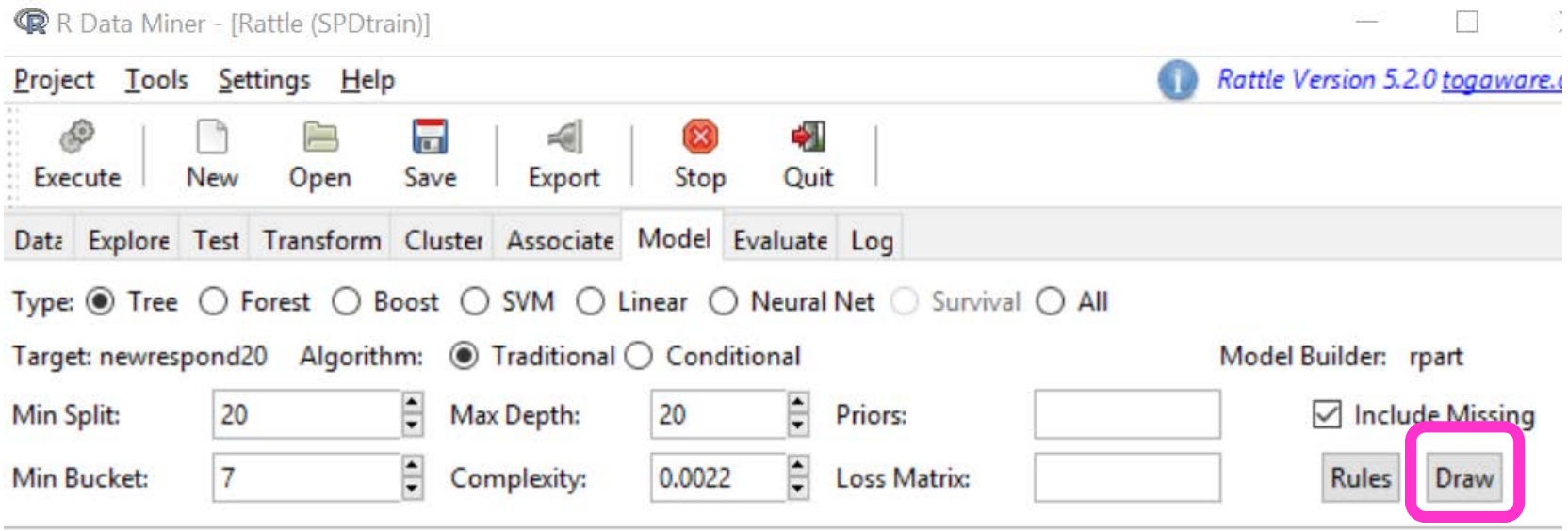
Step 2: Specify the Model

- To start we will build a tree model with a complexity parameter of

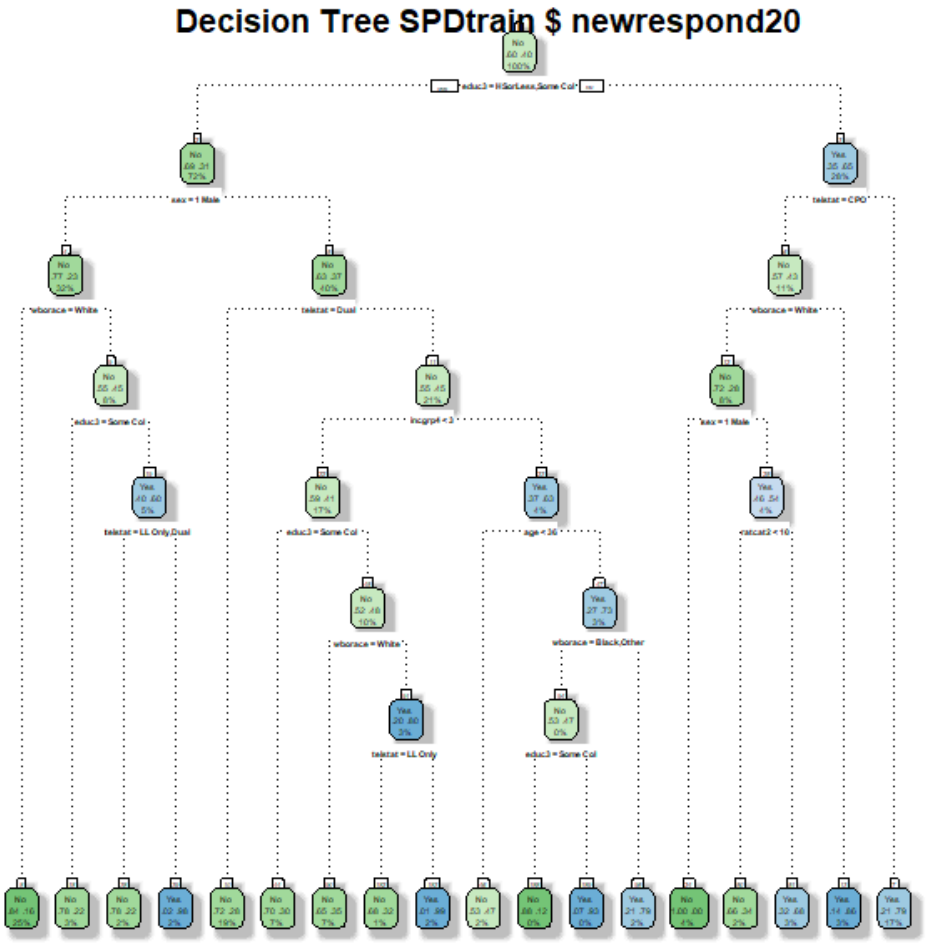




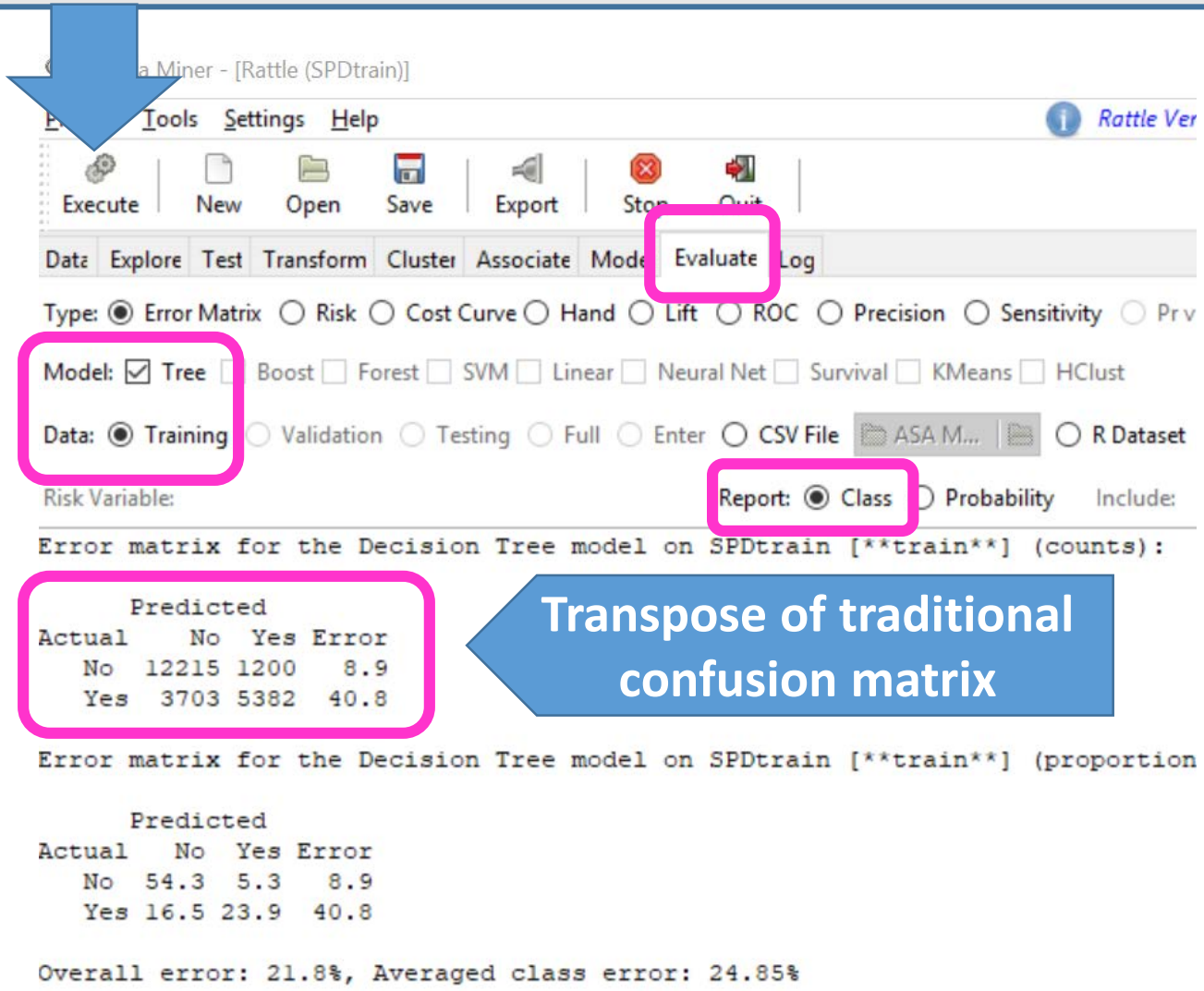
Step 3: Draw the Tree



Tree Depiction



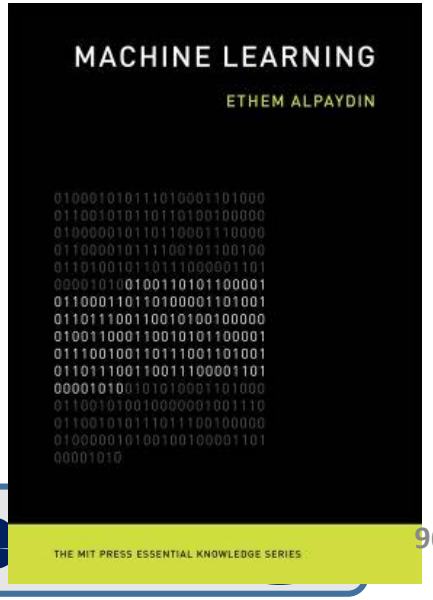
Step 4: Evaluate Model



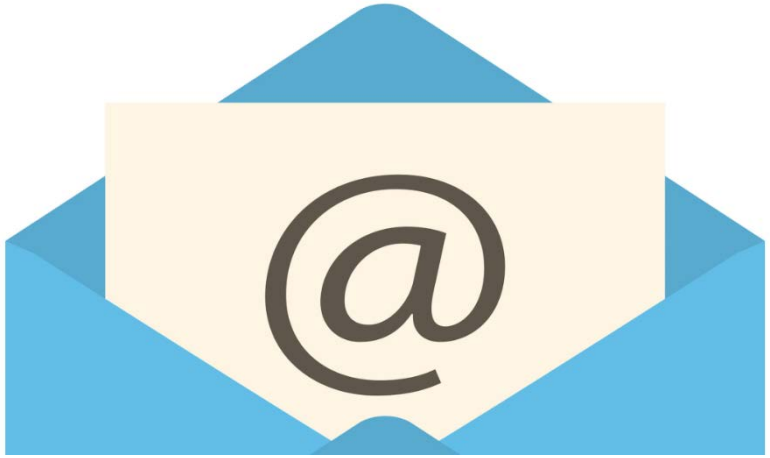
In Closing…

"Machine learning will help us make sense of an increasingly complex world. Already we are exposed to more data than what our sensors can cope with or our brains can process."

Ethem Alpaydin in his book,
Machine Learning: The New AI (2016)



Thank You!



buskirk@bgsu.edu

Bit.ly/BGSUBuskirk

References

- **AAPOR (2015). AAPOR Report on Big Data. Available Here: http://www.aapor.org/AAPORKentico/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15.pdf**
- Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*. Pacific Grove: Wadsworth.
- Buskirk, T.D., West, B. T. and Burks, A -T (2013) "Respondents: Who Art Thou? Comparing Internal, Temporal, and External Validity of Survey Response Propensity Models Based on Random Forests and Logistic Regression Models," Presented at the 2013 Joint Statistical Meetings of the American Statistical Association, Montreal, Canada.
- Buskirk, T.D., Bareham, J.S., Bordy, N. and Dalbey, D. (2014) "Exploring Demographic, Geospatial and Household Correlates of U.S. County-Level Household CPO Rates," Presented at the 2014 MAPOR Annual Conference, Chicago, IL.
- **Buskirk, T. D. & Kolenikov S. (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.* Retrieved from http://surveyinsights.org/?p=5108**
- Earp, M, Mitchell, M., McCarthy, J. and Kreuter, F. *(2014). Modeling Nonresponse in Establishment Surveys: Using an Ensemble Tree Model to Create Nonresponse Propensity Scores and Detect Potential Bias in an Agricultural Survey, Journal of Official Statistics*, Vol. 30(4), 701–719
- Elliott, M. R. (2011). A Simple Method to Generate Equal -Sized Homogenous Strata or Clusters for Population-Based Sampling. *Annals of Epidemiology*,*21*(4), 290–296. doi:10.1016/j.annepidem.2010.11.016
- Krantz, A. Korn, R. and Menninger, M. (2009). Rethinking Museum Visitors: Using K-means Cluster Analysis to Explore a Museum's Audience. *Curator*, Vol. 52 (4), 363-374.
- Barcaroli, G. (2014). Optimization of sampling strata with the SamplingStrata package, available at: ftp://cran.r-project.org/pub/R/web/packages/SamplingStrata/vignettes/SamplingStrataVignette.pdf , accessed on March 21, 2015.
- McCarthy, J.T., Jacob, T. and Atkinson, D. (2009). Innovative Uses of Data Mining Techniques in the Production of Official St atistics. *Federal Committee on Statistical Methodology Papers,* accessed from https://fcsm.sites.usa.gov/files/2014/05/2009FCSM_McCarthy_X-A.pdf , April 1, 2015
- **Mendez, G., Buskirk, T.D., Lohr, S. and Haag, S. (2008) Factors Associated with Persistence in Science and Engineering Majors: An Exploratory Study Using Random Forests. *Journal of Engineering Education*, Vol. 97, No.1, pp. 57-70. See: http://bit.ly/1EULVv4**
- Mesa D.M., Tsai P., Chambers R.L. (2000*), Using Tree-Based Models for Missing Data Imputation: an Evaluation Using UK Census Data*, Report, University of Southampton.
- Molinaro AM, Simon R., Pfeiffer RM. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*. 21(15):3301-7. See http://bit.ly/1bWRCLG
- Ratner, B. (2012) Statistical and Machine-Learning Data Mining, 2nd Ed. CRC Press, Boca Raton.
- Schouten,B., Calinescu, M. and Luiten, A. (2013). Optimizing quality of response through adaptive survey designs. *Survey Methodology*, 39(1), 29-58.
- Wagner, J. and Hubbard, F. (2014). Producing Unbiased Estimates of Propensity Models During Data Collection. *Journal of Survey Statistics and Methodology*, 2(3), 323-342.

Run the KNN Classifier in R: Determine optimal number of Neighbors using 5-fold CV

```
require(survey)
require(caret)
require(gplots)
require(Rcmdr)
#partition the apipop data by school type and select high schools
highS<-split(apipop, apipop$stype)$"H"
PredVars<-names(highS)[c(20,21,23,35)]
#remove 2 NA values
highSc<-highS[complete.cases(highS[,PredVars]),]
Outcome<-names(highSc)[16]

#Now we will create our "training" and "test" sample consisting of 80% and 20% of
#the full dataset.
set.seed(13018)
test<-sample.int(n=dim(highSc)[1], size=150)
trainHS<-highSc[-test,]
testHS<-highSc[test,]
```

Run the KNN Classifier in R: Determine optimal number of Neighbors using 5-fold CV

```
# Run the KNN model with 5-fold cross validation on the training sample
set.seed(2018)
HSctrl<-trainControl(method="cv", number=5, returnResamp="all")
HSknn<-train(trainHS[,PredVars],y=trainHS[,Outcome], method="knn",
tuneGrid=data.frame(.k=c(2*(0:25)+1)),trControl=HSctrl)
# Plot the resulting values of accuracy versus k - cross validated #averages
# Plot not shown, but matches the mean values on plot on next slide
# Means plots representing 5-fold averages/standard deviations for accuracy
plot(HSknn, lwd=3, col=2)
# Accuracy ranges and means plot by k is produced using the following code:
HScvresults<-HSknn$resample
HScvresults<-HScvresults[order(HScvresults[,3], HScvresults[,4]),]
HScvresults <- within(HScvresults, {kfact <- as.factor(k)})
with(HScvresults, plotMeans(Accuracy, kfact, error.bars="se", col=2, lwd=3))
#Determine the CV-Optimal Neighborhood size:
HSknn$bestTune # 25 - optimal # of neighbors for classification based on 5-fold CV
```

Determine the performance of the final model tuned on the Training data and applied to the Test data

```
# We apply the HSknn final model to the test dataset; using the predicted values
# for meeting the API target along with the actual values we know for
# High Schools in our test data we compute the confusion matrix and
# Corresponding accuracy metrics.

HSconfuse<-confusionMatrix(data=HSknnpred25, highSc[test,Outcome],
positive=c("Yes"))
Hsconfuse

# this code will represent the confusion matrix as a heat plot/table.
require(ggplot2)
confusion <- as.data.frame(HSconfuse$table)
plot <- ggplot(confusion)
plot + geom_tile(aes(x=Prediction, y=Reference, fill=150-Freq)) +
scale_x_discrete(name="Predicted Class") + scale_y_discrete(name="Actual
Class") + scale_fill_gradient(breaks=seq(from=-4, to=0, by=1)) +
labs(fill="Normalized \n Frequency")
```