Building Better Models

Introduction to Modeling

Bulding Better Models – Part 1

jmp® Statistical Discovery.™ From SAS.

sas® THE POWER TO KNOW.®

The Goal of Analytics

- Extracting information from data[1] to make actionable, data driven decisions.
- Shmueli[2] breaks analytical goals into three types:
  - Description – Summarizing or representing the data structure in a compact manner. Summary statistics, data visualization, and classification are examples.
  - Explanation – Establishing and quantifying the causal relationships between inputs and outcomes[3]. Testing causal models falls into this group.
  - Prediction – Divination of future outcomes based on current inputs.

1. Hand, D.J. (2008) Statistics: A Very Short Introduction. Oxford University Press, Oxford.
2. Shmueli, G. (2010) To explain or to predict? Statistical Science, 25, pp. 289–310.
3. Kenett, R.S. & Shmueli, G. (2016) Information Quality. The Potential of Data and Analytics to Generate Knowledge. John Wiley & Sons, New York.

Challenges of Big Data

- Variables from historical data are likely to be correlated. Which is the best/smallest set of variables to describe, explain, or predict the questions of interest?
- The bigger the data, the more computational time needed for analysis.
- More data makes it possible to consider more complex relationships. Modeling techniques that can capture these complexities without overfitting are needed.

What is a Model?

$$\mathcal{Y} = \mathcal{F}(\mathcal{X})$$

- Construct $\mathcal{X}$ is related to construct $\mathcal{Y}$ through function $\mathcal{F}$.
- Operationalized into mathematical function:

$$\boldsymbol{Y} = f(\boldsymbol{X}) + \boldsymbol{E}$$

- Where $\boldsymbol{Y}$ and $\boldsymbol{X}$ are quantifiable phenomena. $\boldsymbol{E}$ accounts for the disparity between the fitted model $f(\boldsymbol{X})$ and $\boldsymbol{Y}$.

Shmueli, G. (2010). To Explain or to Predict. *Statistical Science*, 25(**2**), 283 – 310

What is a Model?

$$Y = f(X) + E$$

- An empirical representation that relates a set of inputs (predictors, X) to one or more outcomes (responses, Y)
  - Y is one or more continuous or categorical response outcomes
  - X is one or more continuous or categorical predictors
  - $f(X)$ describes predictable variation in Y (signal)
  - E describes non-predictable variation in Y (noise)
- The mathematical form of $f(\mathbf{X})$ can be based on domain knowledge or mathematical convenience.
- "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."

  – George Box

What makes a Good Model?

- Accurate representation of the population of interest.
- Identification of important predictors. Accurate and precise estimation of predictor effects.
  - Focus is on the right hand side of the above equation.
  - Interest is in the influence predictors have on the response(s).
  - Design of experiments is an important tool for achieving this goal. Good designs lead to good models.
- Accurate and precise prediction.
  - Focus is on the left hand side of the above equation.
  - Interest is in best predicting outcomes from a set of inputs, often from historical or in situ data.
  - Predictive modelling (data mining) methodologies are important tools for achieving this goal.

Additional Considerations

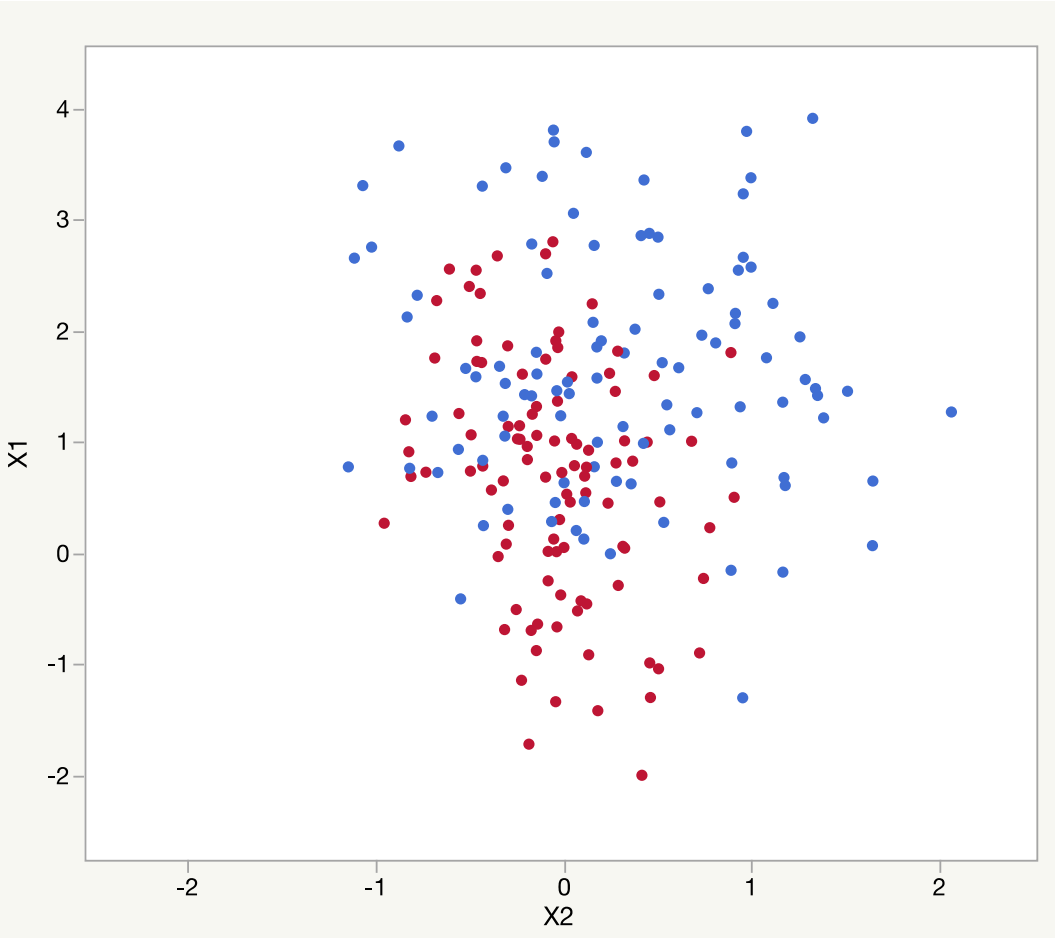
- Are the data representative of the problem being model? Have changes occurred since the last time the problem was modeled? Do I have all the data that I need to adequately model the problem?
- Are there missing values or outliers? Do predictors need to be transformed or imputed? Do I have superfluous predictors? Are the data values correct?
- Is the data set organized in a fashion that is conducive to analysis?
- Have I selected an adequate modeling technique for the underlying construct and the goals of the problem? Does my model adequately reflect the behavior of the construct.

Introduction to Predictive Models

- $f(\mathbf{X})$: A majority of techniques are captured by two general approaches:
  - Global function of data
    - » More stable, less flexible
    - » Captures smooth relationship between continuous inputs and outputs
    - » Examples: Regression, Generalized Regression, Neural Networks, PLS, Discriminant Analysis
  - Local function of data
    - » More flexible, less stable
    - » Captures local relationships in data (e.g., discrete shifts and discontinuities)
    - » Examples: Nearest Neighbors, Bootstrap Forest, Boosted Trees

Introduction to Predictive Models

- Example: Predict group (red or blue) using X1 and X2



Preventing Model Overfitting

- If the model is flexible what guards against overfitting (i.e., producing predictions that are too optimistic)?
  - Put another way, how do we protect from trying to model the noise variability as part of $f(\mathbf{X})$?
- Solution – Hold back part of the data, using it to check against overfitting. Break the data into two or three sets:
  - The model is ***built*** on the training set.
  - The validation set is used to ***select*** model by determining when the model is becoming too complex
  - The test set is often used to ***evaluate*** how well model predicts independent of training and validation sets.
  - Common methods include k-fold, random holdback and bootstrapping.

Predictive Modeling

- Common predictive modeling techniques:
  - Linear regression
  - Generalized regression
    - » Also known as penalized regression or shrinkage methods. It is a technique applied to linear regression to account for correlated inputs. Ridge regression, LASSO, and Elastic Net are three examples.
  - Tree based methods: Partition (CART), Bootstrap Forest (Random Forest), and Boosted Tree.
  - Neural networks
  - Principal components regression (PCR) and partial least squares (PLS).

Regression and Model Selection

Building Better Models – Part 1

jmp® Statistical Discovery.™ From SAS.

sas® THE POWER TO KNOW.®

Regression

- General linear regression typically uses simple polynomial functions for $f(\mathbf{X})$.
  - For continuous y:

$$f(x) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \sum_{i=1}^{p} \sum_{j=i+1}^{p} \gamma_{i,j} x_i x_j + \sum_{i=1}^{p} \delta_i x_i^2$$

  - For categorical y, the logistic function of $f(\mathbf{X})$ is typically used.

$$\frac{1}{1+e^{-f(x)}}$$

Model Selection

- Stepwise Regression
  - Start with a base model: intercept only or all terms.
  - If intercept only, find term not included that explains the most variation and enter it into the model.
  - If all terms, remove the term that explains the least.
  - Continue until a (typically) p-value based stopping criterion is met.
- A variation of stepwise regression is all possible subsets (best subset) regression.
  - Examine all 2, 3, 4, …, etc. term models and pick the best out of each. Sometimes statistical heredity is imposed to make the problem more tractable.

Generalized Regression

- For many engineers and scientists, modeling begins and ends with ordinary least squares (OLS) and stepwise or best subsets regression.
- Unfortunately, when data are highly correlated, OLS may lead to estimates that are less stable with higher prediction variance.
- In addition, there are common situations when OLS assumptions are violated:
  - Errors are not normally distributed.
  - The response is not linear in the parameters.
  - Error variance is not constant across the prediction range.

Penalized Regression

- Generalized regression (aka penalized regression or shrinkage methods) can circumvent these problems.
  - For correlated data, penalized methods produce more stable estimates by biasing the estimates in an effort to reduce prediction estimate variability.
  - Provides a continuous approach to variable selection. More stable than stepwise regression.
  - Can be used for problems when there are more variables then observations, which cannot be estimated using OLS.
- Three common techniques:
  - Ridge Regression: Stabilizes the estimates, but can't be used for variable selection.
  - LASSO
  - Elastic Net: Weighted average of Ridge and LASSO.

Decision Trees

Overview

jmp
Statistical Discovery.™ From SAS.

sas
THE POWER TO KNOW.

Decision Tree Step-by-Step

Goal is to predict those with a code of “1”

Overall Rate is 3.23%

All Rows

Count	G^2
27900	7951.8274

Level	Rate	Prob
0	0.9677	0.9677
1	0.0323	0.0323

Candidates

Term	Candidate G^2	LogWorth
TITLE	2.8610822	1.04217046
CHILDREN	35.8103967	8.75760863
PERS_H	121.2376531	31.99237054
AGE	247.2268705 *	74.63242857
TMADD	32.6734170	7.59168909
TMJOB1	106.3283591	28.30709956
TEL	102.2964249	23.32052298
NMBLOAN	31.7378253	7.86865082
FINLOAN	14.5604894	3.86732124
INCOME	90.5489458	23.86996381
EC_CARD	98.9349359	22.58346708
STATUS	181.8387145	39.50549595
BUREAU	5.5326980	1.48260412
LOCATION	0.2783105	0.22343630
LOANS	9.4887040	1.87838548
REGN	5.0184114	0.83012958
DIV	1.2439822	0.57723792
CASH	19.8373021	4.19284301
PRODUCT	18.7771171	3.21934818
RESID	10.1309318	2.83623745
NAT	14.1823212	1.88384638
PROF	29.8353713	4.91691317
CAR	56.3006567	13.05962720
CARDS	121.3282951	25.68143632

Candidate “X’s”

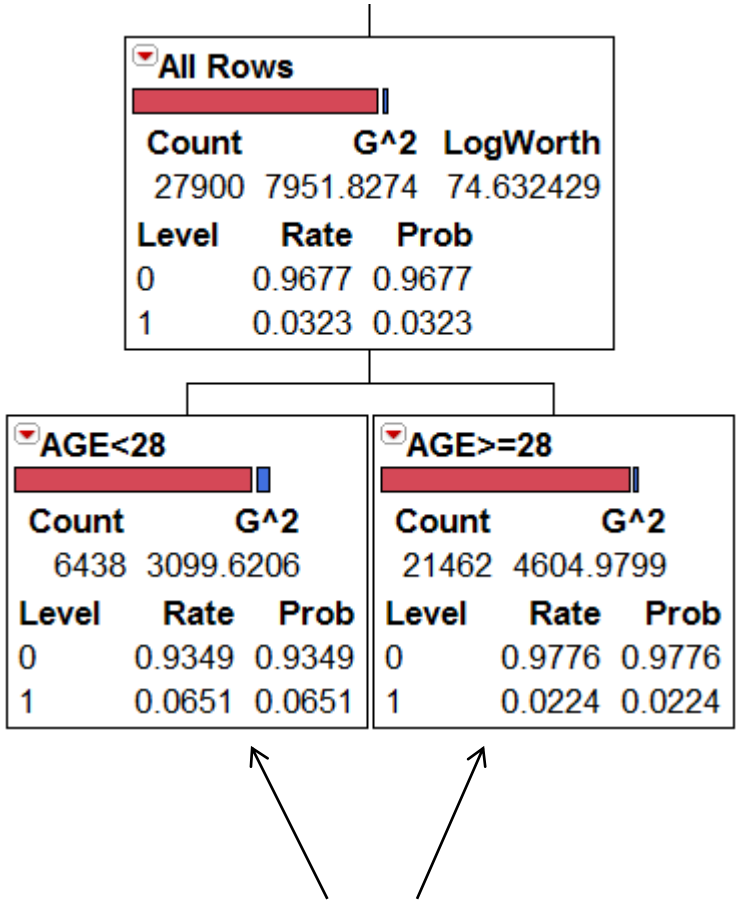
- Search through each of these
- Examine Splits for each unique level in each X
- Find Split that maximizes “LogWorth”
  - Will find split that maximizes difference in proportions of the target variable

Decision Tree Step-by-Step

1st Split:

Optimal Split at Age<28

Notice the difference in the rates in each branch of the tree

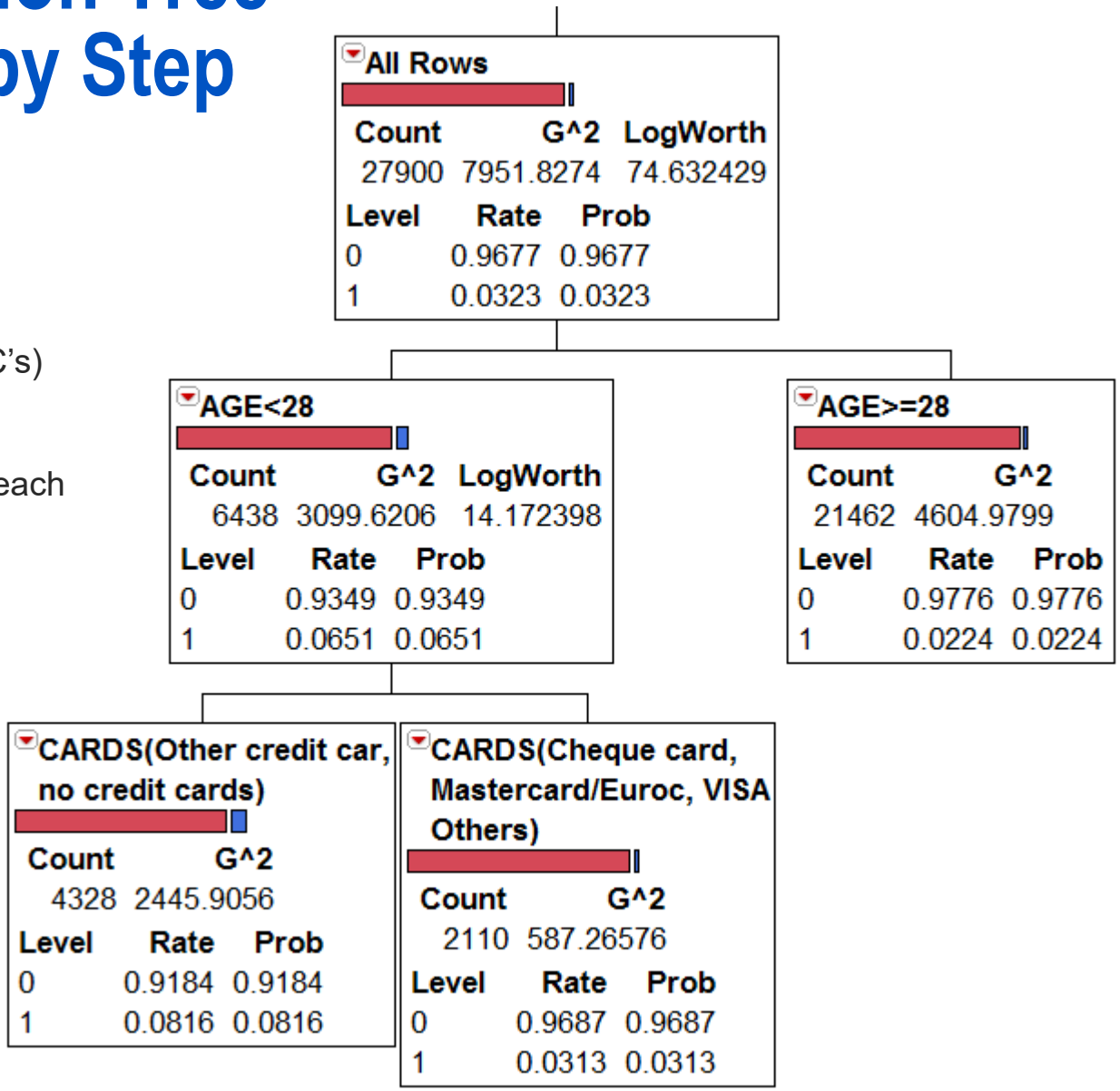


Repeat “Split Search” across both “Partitions” of the data. Find optimal split across both branches.

Decision Tree Step by Step

2nd split on CARDS
(no CC vs some CC's)

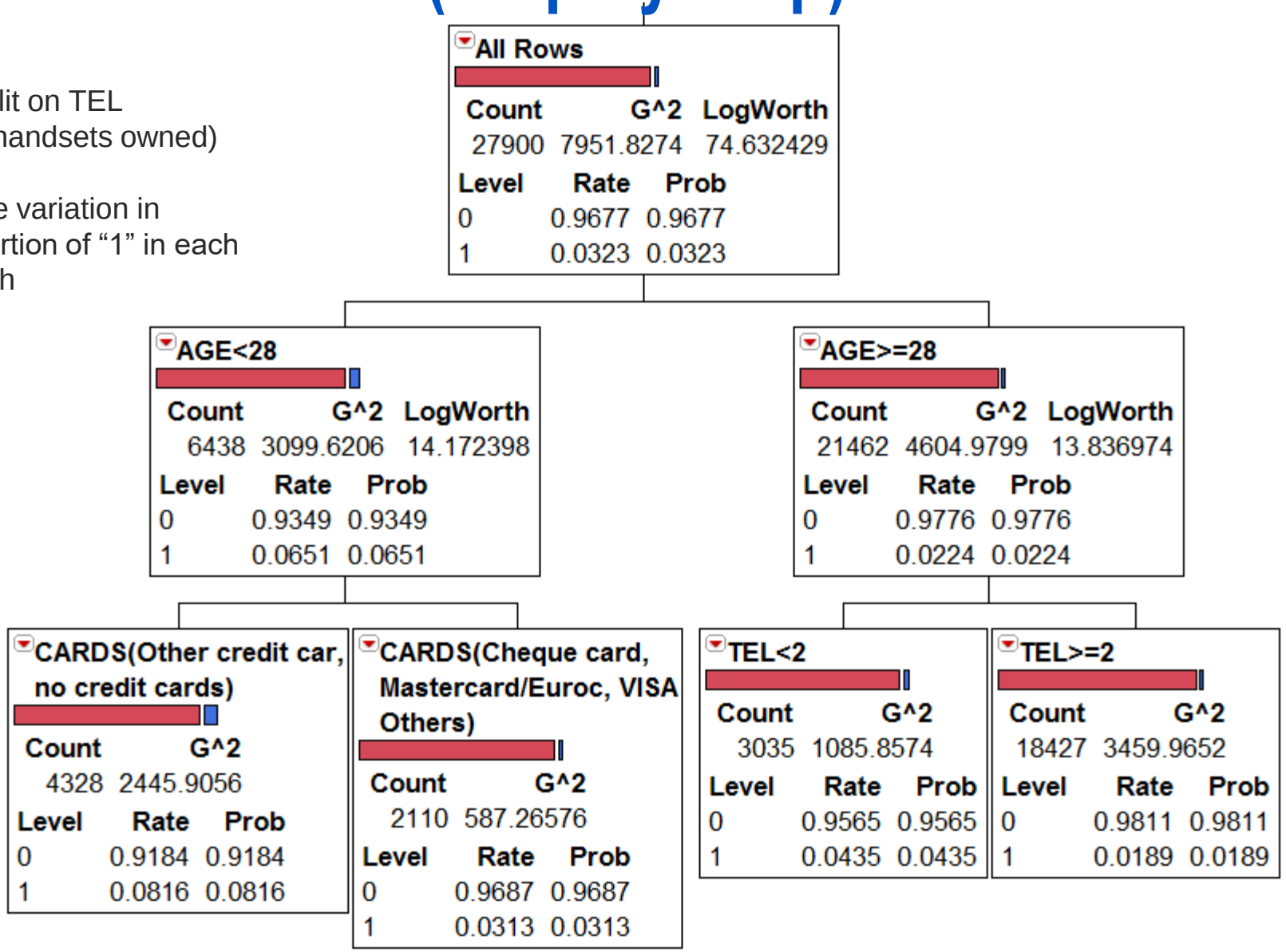
Notice variation in proportion of "1" in each branch



Decision Tree (Step by Step)

3rd split on TEL
(# of handsets owned)

Notice variation in proportion of “1” in each branch

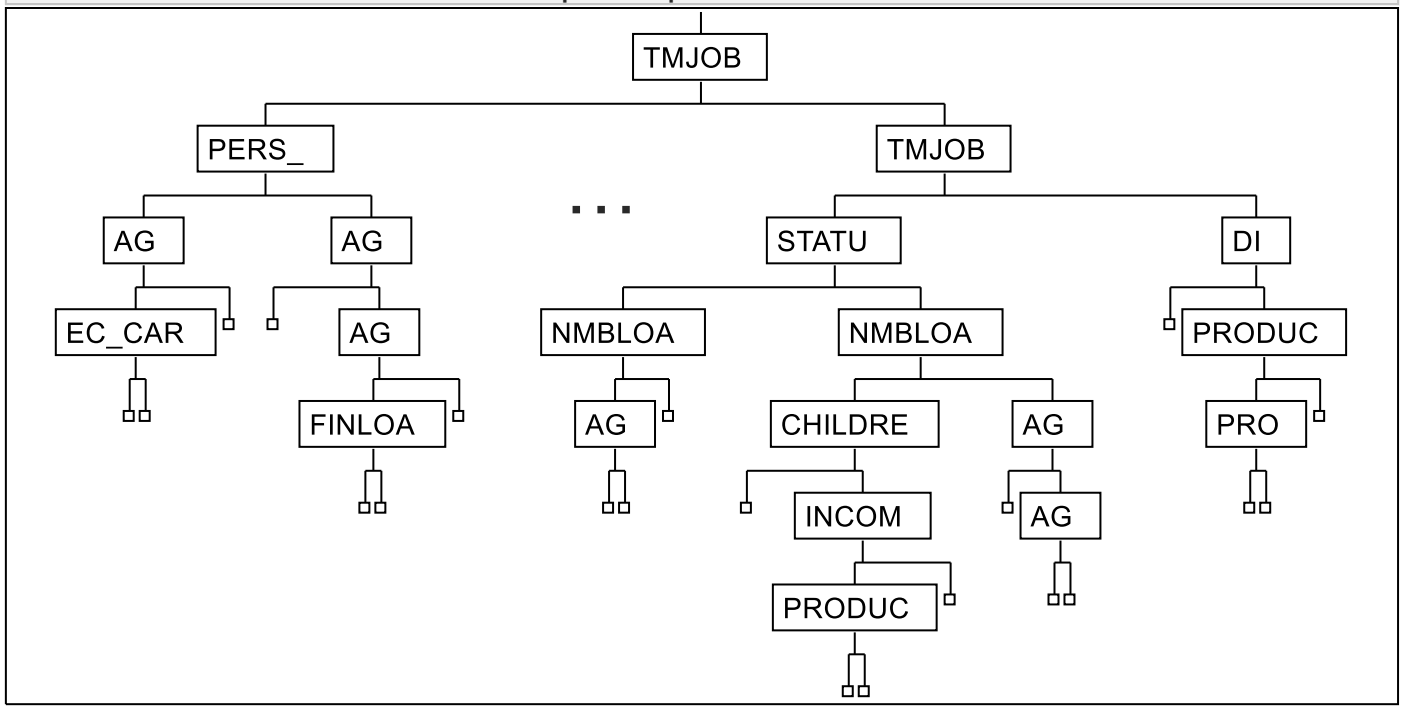


Bootstrap Forest

- Bootstrap Forest
  - For each tree, take a random sample (with replacement) of the data table. Build out a decision tree on that sample.
  - Make many trees and average their predictions (bagging)
  - This is also know as a random forest technique
  - Works very well on wide tables.
- Can be used for both predictive modeling and variable selection.

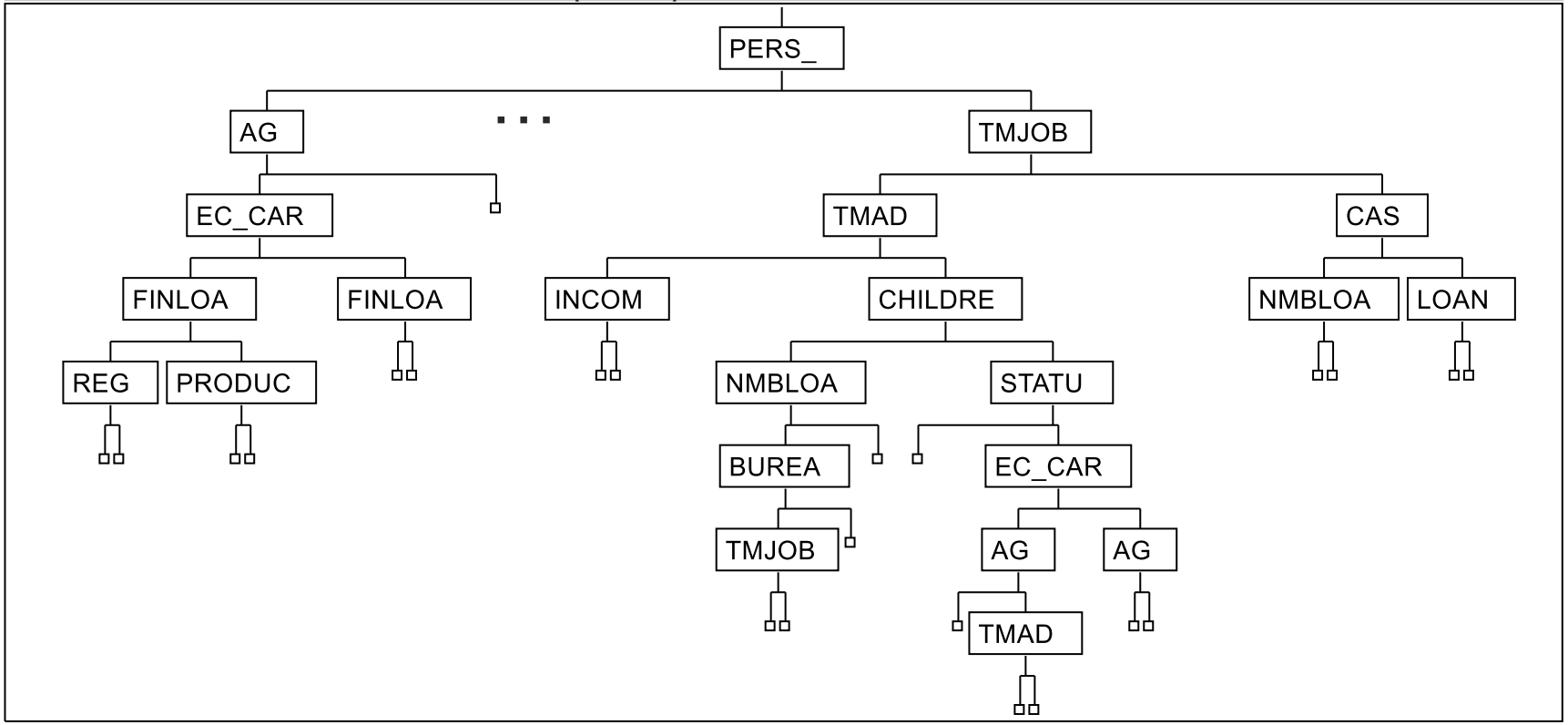
See the Trees in the Forest

Tree1 Tree on 1st Bootstrap Sample



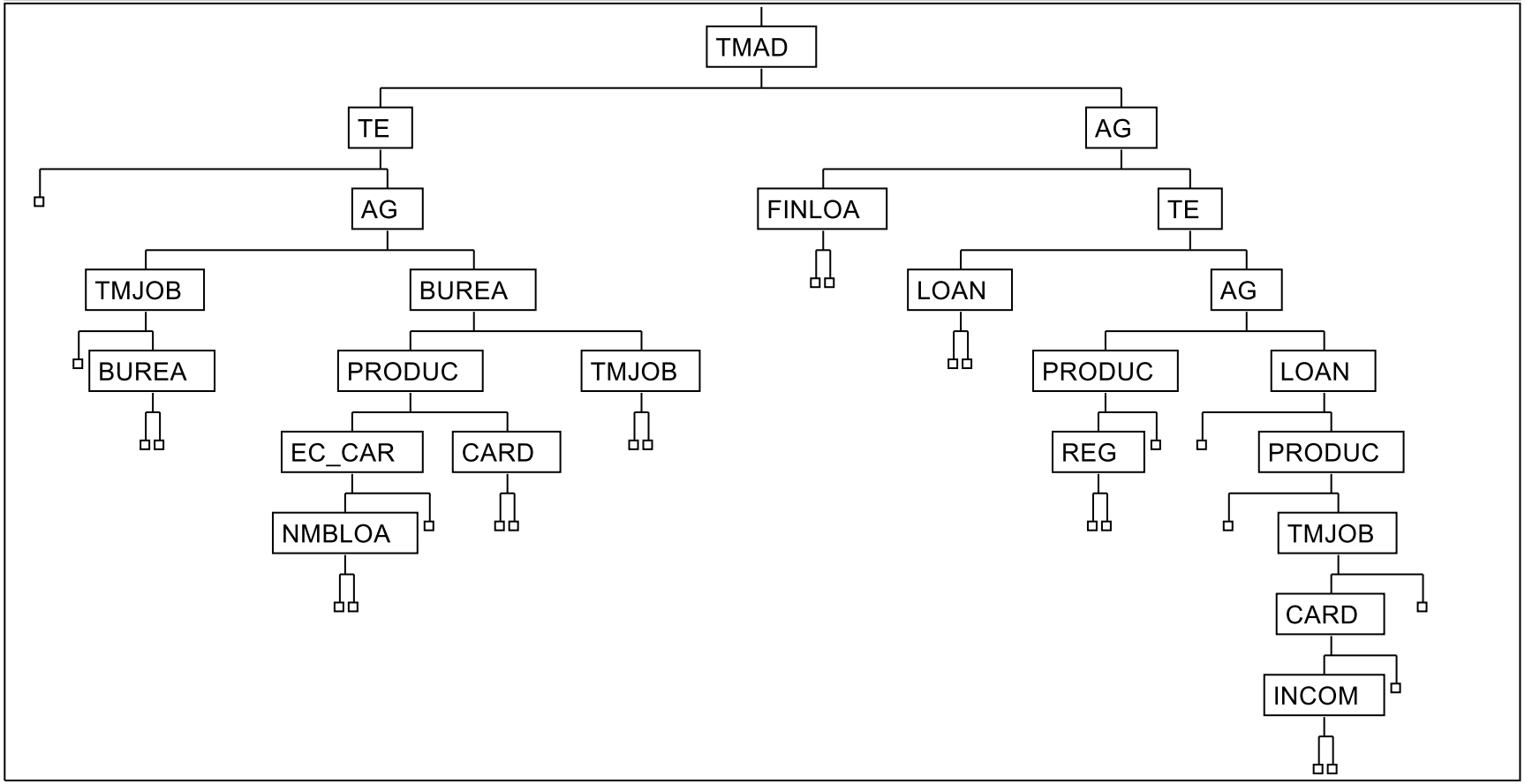
See the Trees in the Forest

Tree2 Tree on 2nd Bootstrap Sample

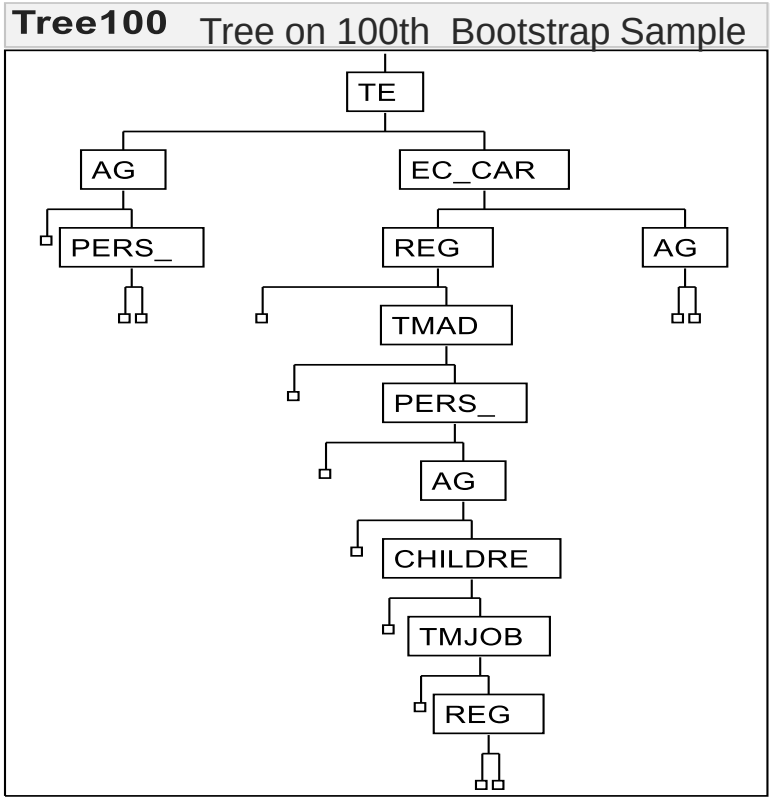


See the Trees in the Forest

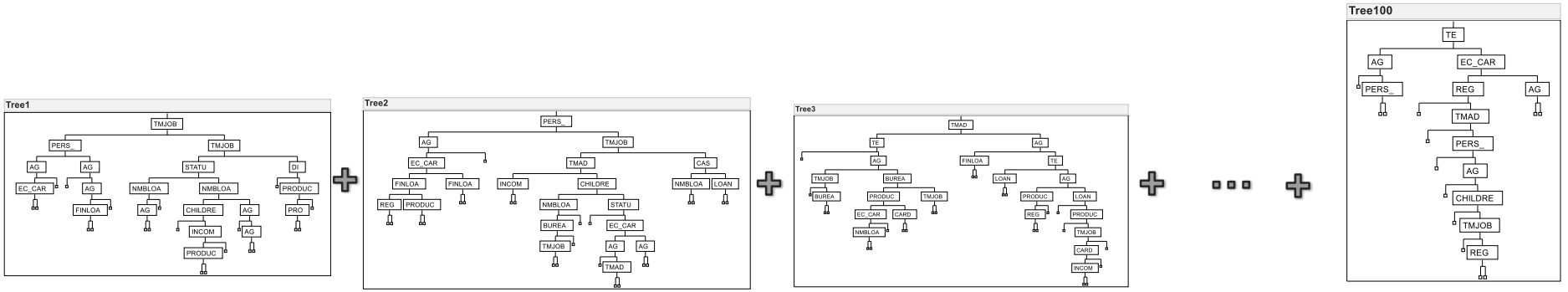
Tree3 Tree on 3rd Bootstrap Sample



See the Trees in the Forest



Average the Trees in the Forest



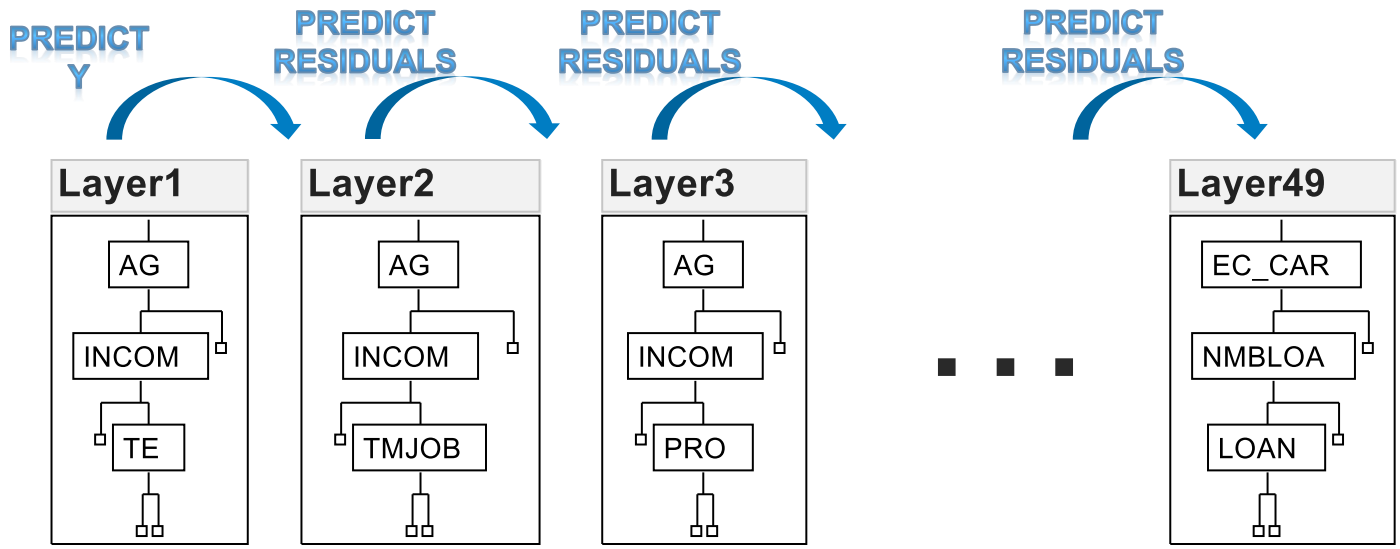
100

Bootstrap Forest Model

Boosted Tree

- Beginning with the first tree (layer) build a small simple tree.
- From the residuals of the first tree, build another small simple tree.
- This continues until a specified number of layers has been fit, or a determination has been made that adding successive layers doesn't improve the fit of the model.
- The final model is the weighted accumulation of all of the model layers.

Boosted Tree Illustrated



Models M1 M2 M3 M49

Final Model

$$M = M1 + \varepsilon \cdot M2 + \varepsilon \cdot M3 + \cdots + \varepsilon \cdot M49$$

ε is the learning rate

Neural Networks

Overview

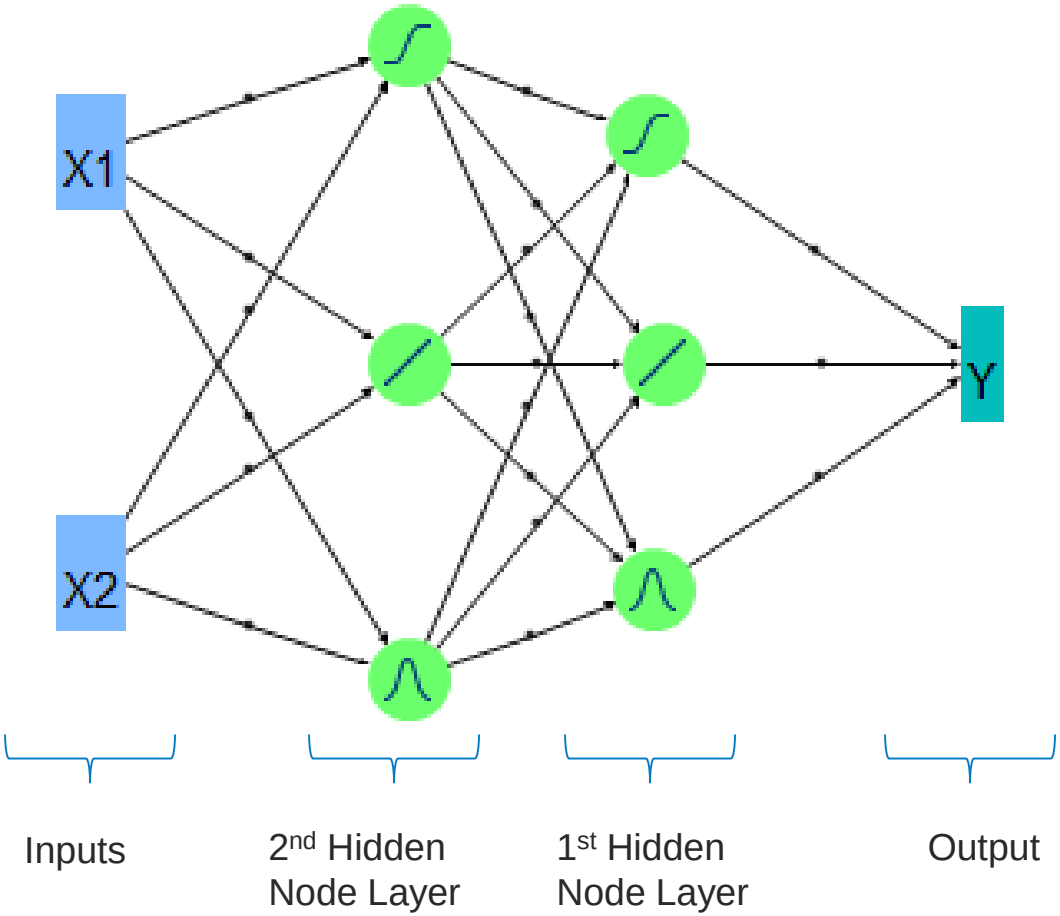
jmp Statistical Discovery.™ From SAS.

sas THE POWER TO KNOW.

Neural Networks

- Neural Networks are highly flexible nonlinear models.
- A neural network can be viewed as a weighted sum of nonlinear functions applied to linear models.
  - The nonlinear functions are called activation functions. Each function is considered a (hidden) node.
  - The nonlinear functions are grouped in layers. There may be more than one layer.
- Consider a generic example where there is a response Y and two predictors $X1$ and $X2$. An example type of neural network that can be fit to this data is given in the diagram that follows

Example Neural Network Diagram



Neural Networks

- Big Picture
  - Can model:
    - » Continuous and categorical predictors
    - » Continuous and categorical responses
    - » Multiple responses (simultaneously)
  - Can be numerically challenging and time consuming to fit
  - NN models are very prone to overfitting if you are not careful
    - » JMP has many ways to help prevent overfitting
      - » Some type of validation is required
      - » Core analytics are designed to stop fitting process early if overfitting is occurring.

See Gotwalt, C., “JMP® 9 Neural Platform Numerics”, Feb 2011, http://www.jmp.com/blind/whitepapers/wp_jmp9_neural_104886.pdf

Model Comparison

Overview

jmp Statistical Discovery.™ From SAS.

sas THE POWER TO KNOW.

Choosing the Best Model

- In many situations you would try many different types of modeling methods
- Even within each modeling method, there are options to create different models
  - In Stepwise, the base/full model specification can be varied
  - In Bootstrap Forest, the number of trees and number of terms sample per split
  - In Boosted Tree, the learning rate, number of layers, and base tree size
  - In Neural, the specification of the model, as well as the use of boosting
- So how can you choose the “best”, most useful model?

The Importance of the Test Set

- One of the most important uses of having a training, validation, AND **test set** is that you can use the test set to assess each model on the same basis.
- Using the test set allows you to compare competing models on the basis of model quality metrics
  - R^2
  - Misclassification Rate
  - ROC and AUC