Interaction Detection with TreeNet®

Dan Steinberg, Mikhail Golovnya, Scott Cardell
Salford Systems
http://www.salford-systems.com

2009

The challenge of interaction detection

- Classical statistical modeling is focused primarily on the development of linear additive modeling
- Models are of the form

$$Y = A + B_1X_1 + B_2X_2 + \ldots + B_kX_k$$

- The predictors or attributes X_i are either raw data columns or data after repairs are made such as missing value imputations and capping or elimination of extreme values
- Non-linearity is introduced into the models in limited ways such as via the log transform of Y (for continuous Y>0) and a collection of well known transforms of the X_i
- Credit risk scorecard technology introduces transforms derived from binning continuous predictors

Classical Statistical Model Performance

- Classical statistical models represent a huge fraction of real world models deployed in enterprises world wide
- Such models are popular in part because statisticians are well trained to develop them
- The models are also popular because they tend to give good performance no matter how such performance is measured (e.g. R-squared, Area Under the ROC curve, top decile lift, etc)
- BUT, such models are so restrictive they must over-simplify
- Nevertheless, they benefit from their stability (low variance)

Bias/Variance Trade-Off

- The universe of possible models is huge whereas the subset of linear or linear models is relatively small
- Forcing a model to conform to a specific mathematical form is likely to introduce distortions
- We are therefore willing to say that the classical models are almost certainly *biased*
  - As data become more plentiful the models almost certainly converge to an incorrect representation of the data generating process
- In contrast, modern learning machines such as TreeNet and CART converge to a "correct" *unbiased* representation
- Classical models have the advantage of *lower variance* which *might* translate into an overall better model in smaller samples

Classical Model Shortcomings: Sources of Bias

- The classical statistical model is expected to fall short in at least two areas:
  - Incorrect representation of nonlinearities in individual predictors
  - Absence of relevant interactions
- Feature selection is also a major problem
- Interactions will be the key to identifying important data segments which behave differently from the dominant patterns
- Specifying a model correctly is the most challenging task facing the modeler
- Until now there has been no reliable way for a modeler to determine whether they have found a correct specification

Classical Interaction Detection

- A popular method of interaction detection is to begin by first building the best possible additive model.
- Then, interactions terms, which look like products of main effects, e.g. X_i*X_j are added to the model and tested for significance
- One problem with this method is that the number of possible interactions grows more rapidly than square of the number of predictors
  - All the usual challenges of model construction apply: the X_i*X_j may not become visible unless one also includes X_m*X_n
  - There may well be higher order interactions $X_i*X_j*X_k$
  - Interactions may only exist and may only be detectable in subregions of the predictors $(X_i|X_i>c)*(X_j|X_j<d)$

What is Needed

- The classical approach of building the best possible additive model and then testing for interactions to add to the model is impractical and largely infeasible
- What is needed is a methodology that can automatically discover the correct transformations of each predictor and introduce the correct interactions
- This methodology must be far more flexible than classical modeling and yet be resistant to overfitting
- It must be able to tell us whether or not interactions are present and to identify precisely which interactions are present

TreeNet and Model Development

- TreeNet is Jerome Friedman's stochastic gradient boosting first described in 1999. Starting from the base of MART (Multiple Additive Regression Trees) TreeNet has evolved over the past decade into a powerful modeling machine.
- TreeNet can fit a variety of models, including regression and logistic regression. The models are non-parametric and based on hundreds if not thousands of small regression trees.
- Each tree is designed to learn very little from the data and thus many are needed to complete the learning process
- Models are in the form of error correcting updates

Some interesting TreeNet Features

- Trees are grown on only a random subset of the training data, typically a random half
  - We never train on all the training data at one time
- Model updates are small. We do not allow the model to change more than a little in any training cycle
- The equivalent of outlying data (points that are badly mispredicted) are eventually ignored in training cycles
  - We do not allow anomalies to have a large influence on the model

A TreeNet Model

- For the binary dependent variable it can be shown that the TreeNet model is a non-parametric logistic regression
- For the continuous dependent variable the TreeNet model is a nonparametric regression fit to maximize (or minimize) one of the following objective functions:
  - Least Squares Residuals
  - Least Absolute Deviations
  - Huber-M hybrid of LS and LAD (LS for small residuals, LAD for large)

Some Key TreeNet Features

- Automatic variable selection. TreeNet is a superb ranking machine.
  - TreeNet tends to include more predictors than other learning machines
  - TreeNet tends to give effective ranking of predictors allowing the modeler to select small subsets of reliable predictors
- TreeNet 2.0 Pro Ex contains built-in step-wise variable selection (backward, forward, or backward/forward) to automatically search for a best model
  - Start with all variables in model
  - Using the variable importance ranking remove the *R* least important predictors (typically *R*=1)
  - Repeat until all predictors have been removed
  - Identify best model based on preferred performance metric

Testing for Interactions

- When the TreeNet model is built with only 2-node trees the model is essentially limited to an additive model
  - With a single split in the tree the tree cannot capture any interactions
  - A 2-node tree TreeNet is thus an additive model.
  - Observe that an additive model can be highly nonlinear
  - Model is of the form
  - $Y = A + F_1(X_1) + F_2(X_2) + \ldots..F_k(X_k)$
  - Where the Fi are the nonlinear functions discovered by TreeNet
- Note that there may be *many Fi* associated with a specific predictor X_j (cumulate these into $G(X_i) = \sum F_q(X_i)$)
- If the TreeNet contains only 2-node trees we can collect all trees associated with a given Xi to arrive at the final model
- $Y = A + G_1(X_1) + G_2(X_2) + \ldots + G_k(X_k)$

Building 2-Node TreeNets

- Important to keep in mind that when only one split is allowed in a tree the amount of learning that can take place is severely limited.
- 2-node tree TreeNets may require a very large number of trees to extract all the information in the data
- Some examples we will show use 20,000 trees
- Unless the TreeNets are fully expanded it will not be possible to measure the true predictive power of the additive model

General TreeNet Models

- TreeNet models are permitted to contain trees of any size
  - We can grow trees with 2,3,4,5,6,12,50…etc nodes
- We generally favor small trees because we want to limit the amount of learning in any training cycle
- We therefore tend to keep trees to sizes like 6,9,12. However experimentation is always recommended and we have encountered data sets where larger trees perform better on holdout data

The Default 6-Node Tree

- Friedman recommended a default setting of 6-nodes and our experiments confirm that this size of tree performs well across broad range of data sets
- A 6-node tree clearly permits interactions. A tree with 6 terminal nodes contains 5 internal splits and if it is as close to balanced as possible like the tree below then it can contain up to 3 different variables along its longest branch
- If the tree is maximally unbalanced then it can contain up to 5 different variables along the longest branch
- There is no guarantee that different variables will be used as we progress down a tree. The same variable might be used several times. In general we observe that the 6 node tree should be adequate to uncover 3-way interactions

Global Interaction Test

- Compare
  - Unrestricted TreeNet model allowing moderate sized trees
  - Restricted TreeNet model confined to just 2-node trees
- Must take into account the *total learning* in each model (number of trees * number of nodes per tree)
  - Otherwise the 2-node tree model will be at an automatic disadvantage
  - Simply allow each model to reach “convergence”

A Global Interaction Test

- Consider a TreeNet model built with 2-nodes and compare this model with a 6-node TreeNet model
- Recall that the TreeNets must have been allowed to grow out fully to locate the optimal number of trees for prediction
- By comparing the restricted 2-node tree model with the 6-node tree model we can conduct a definitive test for the presence of interactions of moderate degree:
  - If the larger tree (unrestricted) model sufficiently outperforms of the (restricted) 2-node model then we have compelling evidence that interactions are present in the data generation process
  - At present we do not have a definitive statistic for testing this hypothesis but classical statistical tests can be developed when comparing predictions of constrained and unconstrained models on holdout data

Some suggested tests

- We can argue for a log-likelihood ratio test; the test statistic will follow a Chi-squared distribution with 1 degree of freedom
- An alternative test:
  - Extract residuals from the 2-node TreeNet (R_TN_N2)
  - Model R_TN_N2 using all original predictors and unrestricted trees
  - If an additive model is sufficient then we should be able to predict these residuals with unrestricted (or less restricted) trees

Which Interactions Are Present

- Our global test is sufficient to establish the existence of interactions but not sufficient to identify which specific interactions are present
- All we can conclude from a positive result of our test is that interactions exist and from a negative result that there is no evidence in favor of interactions
- However this is a major step forward as this test is fully automatic
  - We have found evidence that many mainstream consumer risk models are adequately modeled with additive TNs. This is both a surprise and an important finding.
- The next step for us is therefore to try to identify specific interactions

Interaction Detection In TreeNet

- Interaction in TreeNet has progressed along two fronts. Friedman suggested one strategy in his paper. At Salford Systems, Cardell, Golovnya, and Steinberg suggested a slightly different method.
- In this paper we follow the Salford methodology

Interaction Measurement

- From the TreeNet model extract the function (or smooth)

  $$Y(Xi, Xj \mid Z) \qquad (1)$$

  which is based on averaging the Y associated with all observed Xi, Xj pairs over all observed Z

- Now repeat the process for the single dependencies

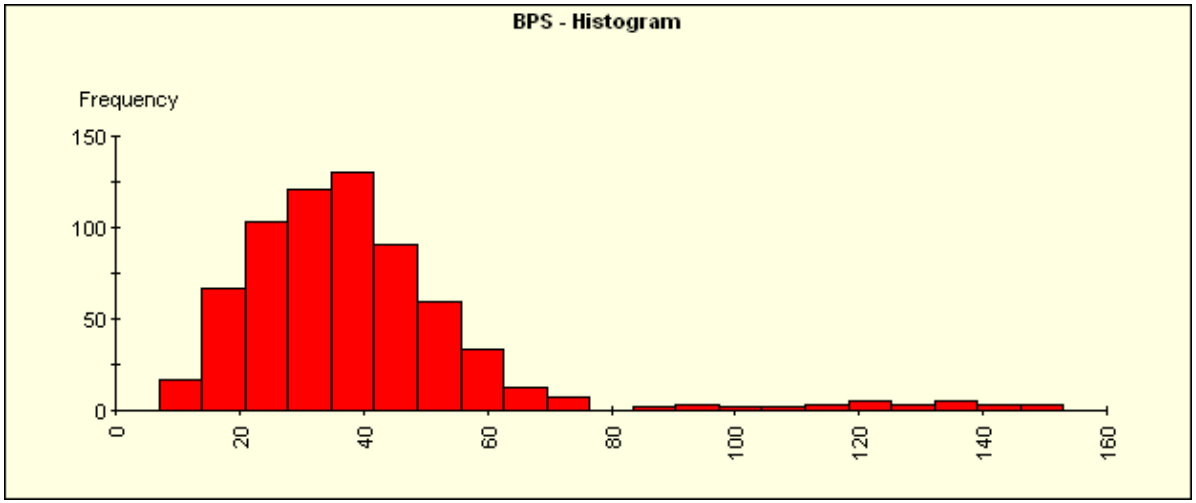
  $$Y(Xi \mid Xj, Z) \qquad (2)$$
  $$Y(Xj \mid Xi, Z) \qquad (3)$$

- Compare the predictions derived from (1) with those derived from (2) and (3) across an appropriate region of the (Xi, Xj) space
- Construct a measure of the difference between the two 3D surfaces (1) and (2)+(3)

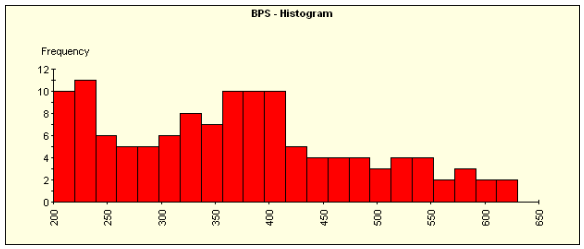
An illustrative example

- In the next slides we walk the process of interaction detection and verification using a small real world data set
- The example concerns regression, the prediction of a continuous dependent variable, in this case the value of a financial security at a point in time
- The distribution of the target variable and summary statistics for the predictors follow

A Simple Example: Financial Market Behavior



Bottom 85% of values



Top 15% of values

Dependent Variable: Continuous with a long right hand tail

Summary Stats: Target and 9 Predictors

Variable	N	N Missing	% Missing	N Distinct	Mean	Min	Max
AVG_DLY_VOL	840	0	0	5	2.7619	1	5
AVG_SIZE	840	0	0	14	62.958	1	873.5
AVG_TRND	840	0	0	5	2.5	1	5
AVG_VOL	840	0	0	14	8.8838e+005	1	16626600
BPS	840	0	0	840	95.744	9.3285	626.4
LIQUIDITY_Q	840	0	0	4	2.0952	1	4
MKT_CAP	840	0	0	4	1.9524	1	4
MOMENTUM_Q	840	0	0	4	2.119	1	4
SPREAD_Q	840	0	0	14	2.3043	-4.4913	6.7593
VOL_Q	840	0	0	4	2.5476	1	4

TARGET variable BPS is truly continuous; others are ideal for a tree model
We randomly set aside 20% of the data for test

Baseline CART Model:
Test Data MSE (Random 20%)

- 110 node tree SE 0 611.86
- 13 node tree SE 1 826.18
- Linear Regression 4639.01

- CART Much better than an off the shelf multiple regression

- Observe that a CART model allows for high order interactions

Naïve Regression: R2 is only .74 on train data

Least Squares Regression: Raw Training Data

N: 658.00 R-SQUARED: 0.7393201
MEAN DEP VAR: 96.4091138 ADJ R-SQUARED: 0.7356996
UNCENTERED R-SQUARED = R-0 SQUARED: 0.8337070

PARAMETER	ESTIMATE	S.E.	T-RATIO	P-VALUE
Constant	97.8222440	17.7186289	5.5208698	0.0000000
MKT_CAP	19.1213408	6.3103432	3.0301586	0.0025416
AVG_DLY_VOL	2.0487381	2.0641444	0.9925362	0.3213065
AVG_TRND	36.2045832	2.5788865	14.0388433	0.0000000
AVG_SIZE	0.2719644	0.0175106	15.5314121	0.0000000
AVG_VOL	0.0000233	0.0000010	23.4757346	0.0000000
MOMENTUM_Q	-7.6325962	2.4448315	-3.1219313	0.0018767
VOL_Q	-32.9141515	2.7359636	-12.0301860	0.0000000
LIQUIDITY_Q	-27.4288542	2.9592253	-9.2689307	0.0000000
SPREAD_Q	-6.5512532	1.4176854	-4.6210910	0.0000046

F-STATISTIC = 204.2008550 S.E. OF REGRESSION = 65.8382733
P-VALUE = 0.0000000 RESIDUAL SUM OF SQUARES = .280887E+07
[MDF,NDF] = [9, 648] REGRESSION SUM OF SQUARES = .796630E+07

Linear regression R2 on test data is .697

MARS Test MSE Models results

- Main Effects (nonlinear) 641.58
- 2-way interactions 190.34
- 3-way interactions 215.11
- 4-way interactions 231.48
- Allowing higher order interactions into MARS here induces some overfitting in this small sample

Cost of Omission	No of Basis Functions	No of Effective Parameters	Variables
4,008.8315796	2	6.222	AVG_SIZE
1,427.4509192	2	6.222	MKT_CAP
1,570.4134311	3	9.333	AVG_SIZE, AVG_VOL
814.7715497	3	9.333	AVG_TRND, AVG_SIZE
1,382.3172949	1	3.111	MKT_CAP, AVG_SIZE
560.8660303	1	3.111	AVG_DLY_VOL, AVG_SIZE
518.4966923	2	6.222	AVG_SIZE, LIQUIDITY_Q
479.7389039	2	6.222	AVG_SIZE, MOMENTUM_Q
657.0530083	4	12.444	AVG_TRND, AVG_SIZE, AVG_VOL
478.7167634	1	3.111	AVG_SIZE, MOMENTUM_Q, LIQUIDITY_Q
507.0676934	1	3.111	MKT_CAP, AVG_TRND, AVG_SIZE, AVG_VOL

MARS main effects and interactions determined by 4-way model

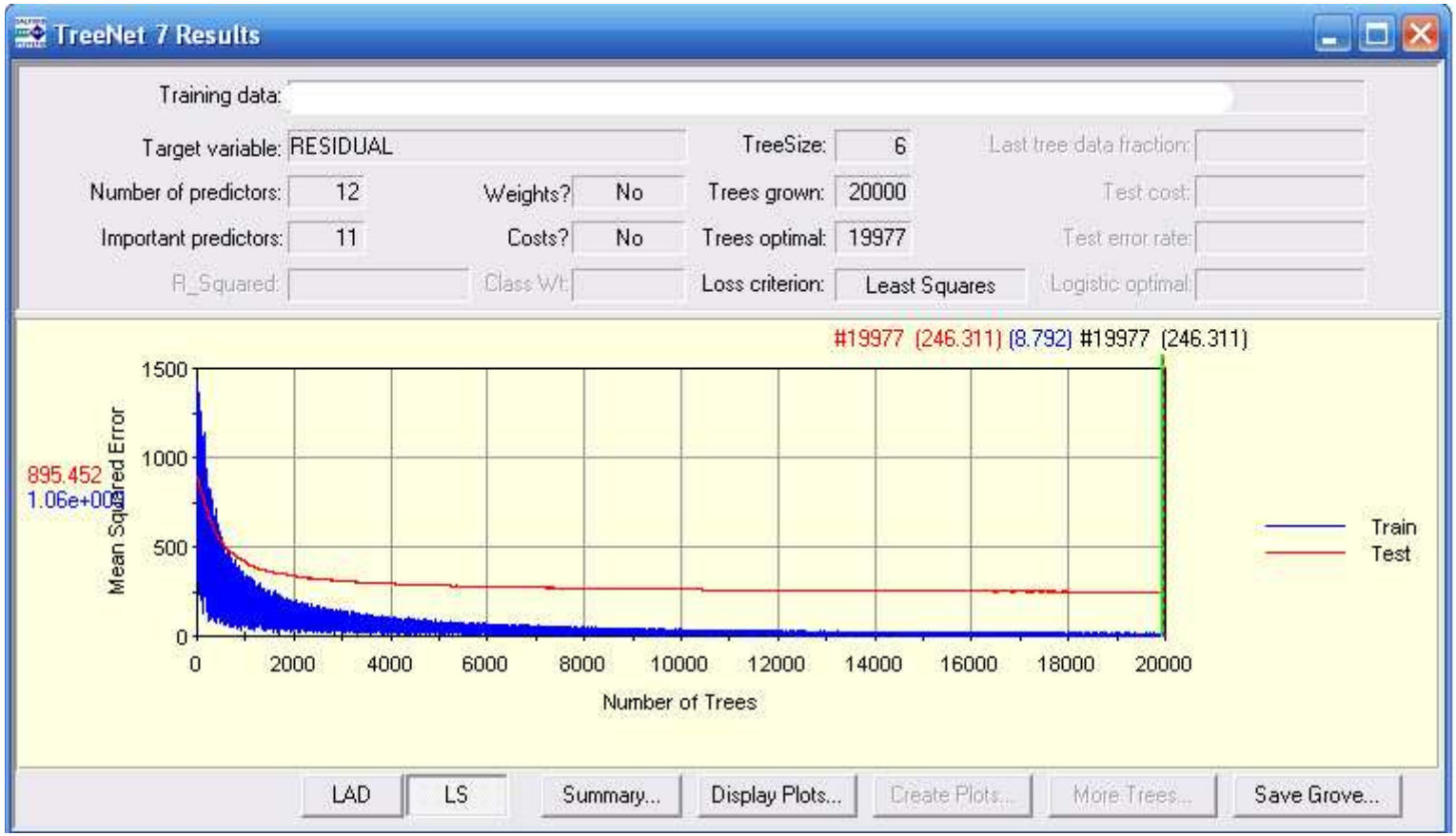
TreeNet Results: Test MSE

- TreeNet 6-node unconstrained (3303 trees) 156.17
- TreeNet 9-node unconstrained (5802 trees) 148.38
- TreeNet 2-node tree (20,000 trees) 891.16

- The 2-node TreeNet underperforms a single CART tree with high order interactions

- The 6-node TN clearly dominates in performance
- Performance can be improved a little by going to a 9-node tree. Larger sizes yield worse performance

Results from the TN Residual Test



Clear signal extracted by the 6-node tree. Residuals are predictable. Test R2=.30

Conclusions From the Global Test

- The results make clear that interactions are critical in this data and ignoring them hurts model performance substantially
- However, from this result we can only conclude that interactions exist
- When we first released TreeNet 1.0 in 2002 we suggested that users review the 3D plots produced by TreeNet to visually search for evidence of strong interactions following the results of 2-node vs 6-node tree results
  - Clearly we want an automated and less error-prone method
- Now we produce detailed interaction reports to eliminate the need for visual inspection

TreeNet Interactions Ranking Report

```
====================
Treenet Interactions
====================

     Spurious interactions tau: 1.00
   # of top 2-way interactions:    5

Whole Variable Interaction Stats
       Abs       Rel Predictor
-----------------------------------------
     31.68    100.00 AVG_VOL
     22.92     72.36 AVG_TRND
     19.32     60.98 SPREAD_Q
      9.43     29.76 MKT_CAP
      7.18     22.65 AVG_SIZE
      2.28      7.19 AVG_DLY_VOL
      1.23      3.87 LIQUIDITY_Q
      1.15      3.63 MOMENTUM_Q
      0.97      3.07 VOL_Q
```

Based on the 6-node tree TreeNet we calculate the degree of interaction observed for the most important variables

This report reveals that e.g. AVG_VOL is involved in important interactions

Detailed Interaction Reports

Predictor: AVG_VOL

Measure1	Measure2	Rel	Predictor
17.74	20.85	100.00	AVG_TRND
17.33	20.24	97.07	SPREAD_Q
4.82	6.37	30.54	MKT_CAP
1.76	2.20	10.53	AVG_SIZE
1.69	2.18	10.44	AVG_DLY_VOL
0.50	0.63	3.03	LIQUIDITY_Q
0.45	0.58	2.77	MOMENTUM_Q
0.22	0.28	1.35	VOL_Q

Predictor: SPREAD_Q

Measure1	Measure2	Rel	Predictor
17.33	20.24	100.00	AVG_VOL
2.46	14.02	69.28	AVG_TRND
2.40	17.78	87.84	AVG_SIZE
0.96	8.01	39.55	LIQUIDITY_Q
0.58	4.71	23.25	MKT_CAP
0.46	4.02	19.85	VOL_Q
0.34	2.69	13.28	AVG_DLY_VOL
0.21	1.83	9.04	MOMENTUM_Q

Predictor: AVG_TRND

Measure1	Measure2	Rel	Predictor
17.74	20.85	84.38	AVG_VOL
4.18	24.71	100.00	AVG_SIZE
2.46	14.02	56.74	SPREAD_Q
0.89	6.17	24.97	MKT_CAP
0.40	2.63	10.63	MOMENTUM_Q
0.37	2.39	9.67	AVG_DLY_VOL
0.26	1.78	7.21	LIQUIDITY_Q
0.22	1.46	5.91	VOL_Q

Predictor: AVG_SIZE

Measure1	Measure2	Rel	Predictor
4.18	24.71	100.00	AVG_TRND
2.40	17.78	71.94	SPREAD_Q
1.76	2.20	8.88	AVG_VOL
1.20	22.04	89.19	MKT_CAP
0.36	7.21	29.18	LIQUIDITY_Q
0.28	6.58	26.62	MOMENTUM_Q
0.18	2.92	11.83	AVG_DLY_VOL
0.11	2.86	11.57	VOL_Q

Measured interactions based on the comparison of 2D and 3D relationships between target and predictors

Allows us to *hypothesize* which interactions are likely to matter

Nine predictors permit a total of 36 two-way interactions

- TreeNet report suggest that the 6 interactions below may be the most important
- Top rated 2-way interactions (6 in all)
  - AVG_VOL * AVG_TREND
  - AVG_VOL * SPREAD_Q
  - SPREAD_Q * AVG_TREND
  - SPREAD_Q * AVG_SIZE
  - AVG_TREND * AVG_SIZE
  - AVG_SIZE * MKT_CAP

Testing Interactions

- Salford Systems has developed an Interaction Control Language (ICL) for TreeNet models
- The language allows the modeler to specify precisely the types of interactions which will be permitted in the TreeNet
- ICL allow X_1 X_2 X_3 X_4 / 2
- Specifies that only 2-way interactions are allowed among the collection of predictors listed (X_1-X_4)
- The ICL language was developed in-house for private clients in 2006 and has been the basis of all of our interaction detection work

The ICL Language

- The ICL allows a broad range range of controls such as:
  - ICL ADDITIVE x_1 x_2 x_3
  - ICL ALLOW x_5 x_6 x_7 x_8 x_9 / 3
  - ICL DISALLOW x_9 x_{11} x_5 x_7 x_{21} x_{25} / 4
- The ADDITIVE keyword prevents any predictor from interacting with any other variable in the model.
  - Practically this means that should such a predictor be selected to split the root node of the tree than it can be the only predictor anywhere in that tree
- TreeNets restricted to ADDITIVE predictors can still contain trees with as many terminal nodes as the modeler prefers to work with.
  - But each tree will be grown using a single predictor

Test MSE Performance: Restricted Models

- 2-node trees (20,000 trees) 891.16

Whole Variable Interaction Stats

Abs	Rel	Predictor
0.00	100.00	AVG_VOL
0.00	99.94	VOL_Q
0.00	99.84	SPREAD_Q
0.00	99.72	MKT_CAP
0.00	99.70	LIQUIDITY_Q
0.00	99.67	MOMENTUM_Q
0.00	99.66	AVG_SIZE
0.00	99.52	AVG_DLY_VOL
0.00	99.28	AVG_TRND

- As expected the interaction report shows 0.00 for all interaction scores. The data contain no missing values so literal 2-node trees are generated for this model
- We require 20,000 trees because 2-node trees can learn only little in any training cycle

Test MSE Performance: Various Interactions Allowed

- 2-node trees (20,000 trees) 891.16
- Allow 2-way interactions 163.31
- Allow 3-way interactions 158.35
- Allow 4-way interactions 157.49
- Allow 5-way interactions 158.49
- Unconstrained TreeNet 156.17

- It is plain that there is a huge difference between an additive and an interaction model. Much less difference between models with differing degree of interaction allowed

Refining the Model

- Using the ICL mechanism we constrain the TreeNet

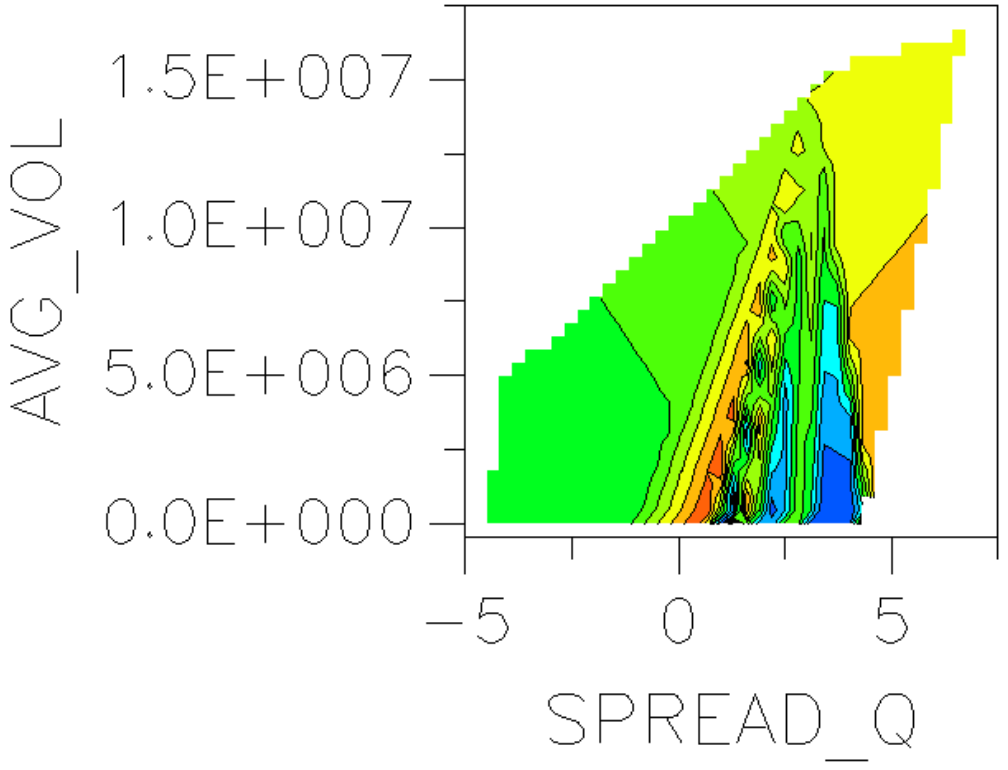
Model	Test MSE
Unrestricted TreeNet	156.17
Allow all predictors in 2-way interactions	163.31
Allow 8 predictors in 2-way interactions	166.34
Allow 7 predictors in 2-way interactions	166.94
Allow 6 predictors in 2-way interactions	167.16
Allow 5 predictors in 2-way interactions	165.55
Allow 4 predictors in 2-way interactions	170.25
Allow 3 predictors in 2-way interactions	175.02
Allow 2 predictors in 2-way interactions	171.90

Select Two Way Interactions

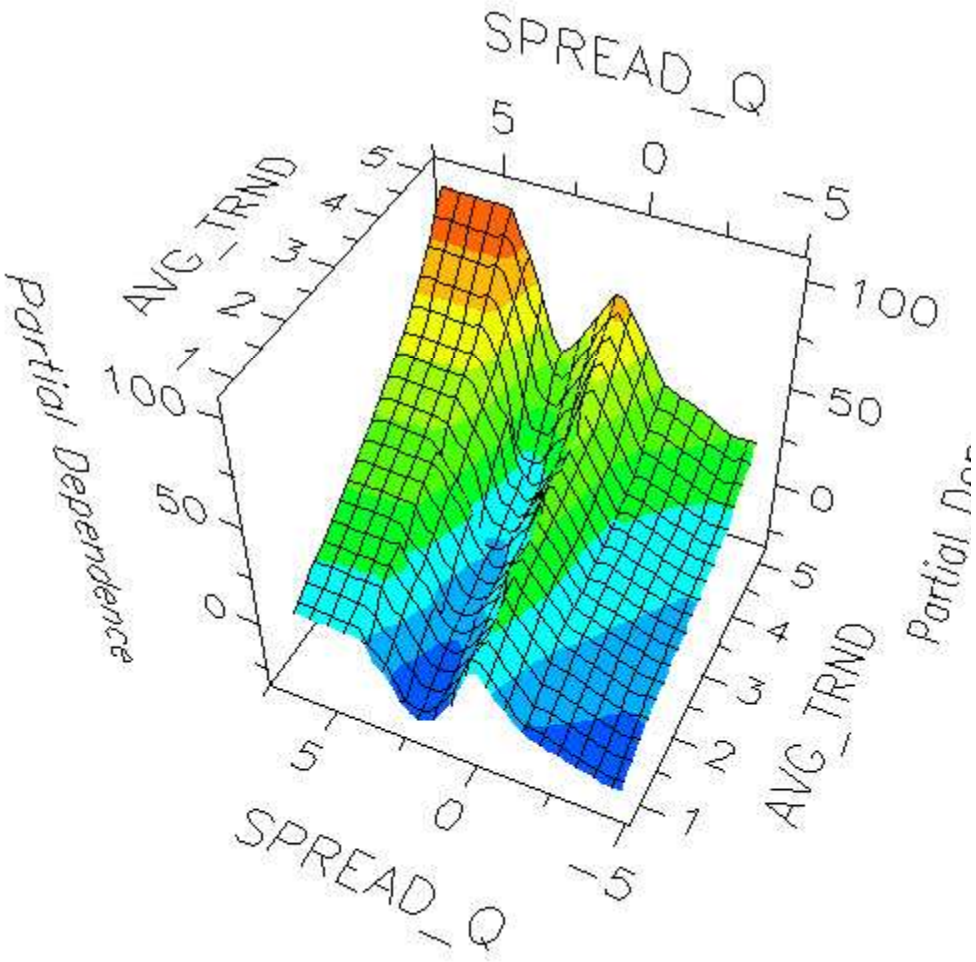
- The following slides show 3D graphs for some powerful interactions extracted from the TreeNet output

AVG_VOL * SPREAD_Q

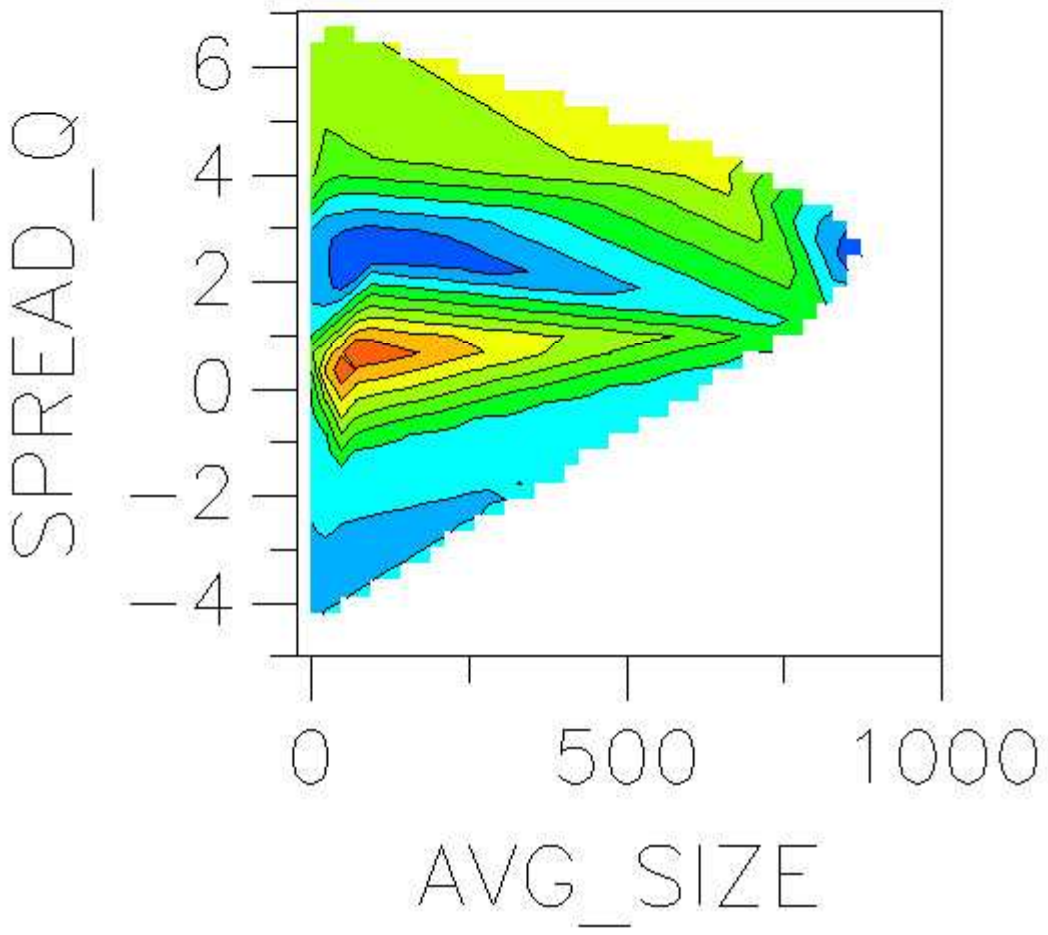
- For most values of SPREAD_Q the marginal impact on BPS varies strongly by the value of AVG_VOL



SPREAD_Q * AVG_TRND



SPREAD_Q* AVG_SIZE



TreeNet with ICL

- TreeNet in the PRO EX version contains the ICL language and is available on request from Salford Systems.
- Contact:
  - David Tolliver
  - dst@salford-systems.com
  - 619 543 8880
- General information
  - http://www.salford-systems.com

ADDITIVE vs 2-node Trees

- 2-node trees are often thought to guarantee ADDITIVE models but this is not strictly true
- If the training data are complete (no missing values present) then indeed a 2-node tree TreeNet yields an additive (in the predictors) model
- However, if missing values are present then TreeNet requires the use of missing value indicator predictors of the form

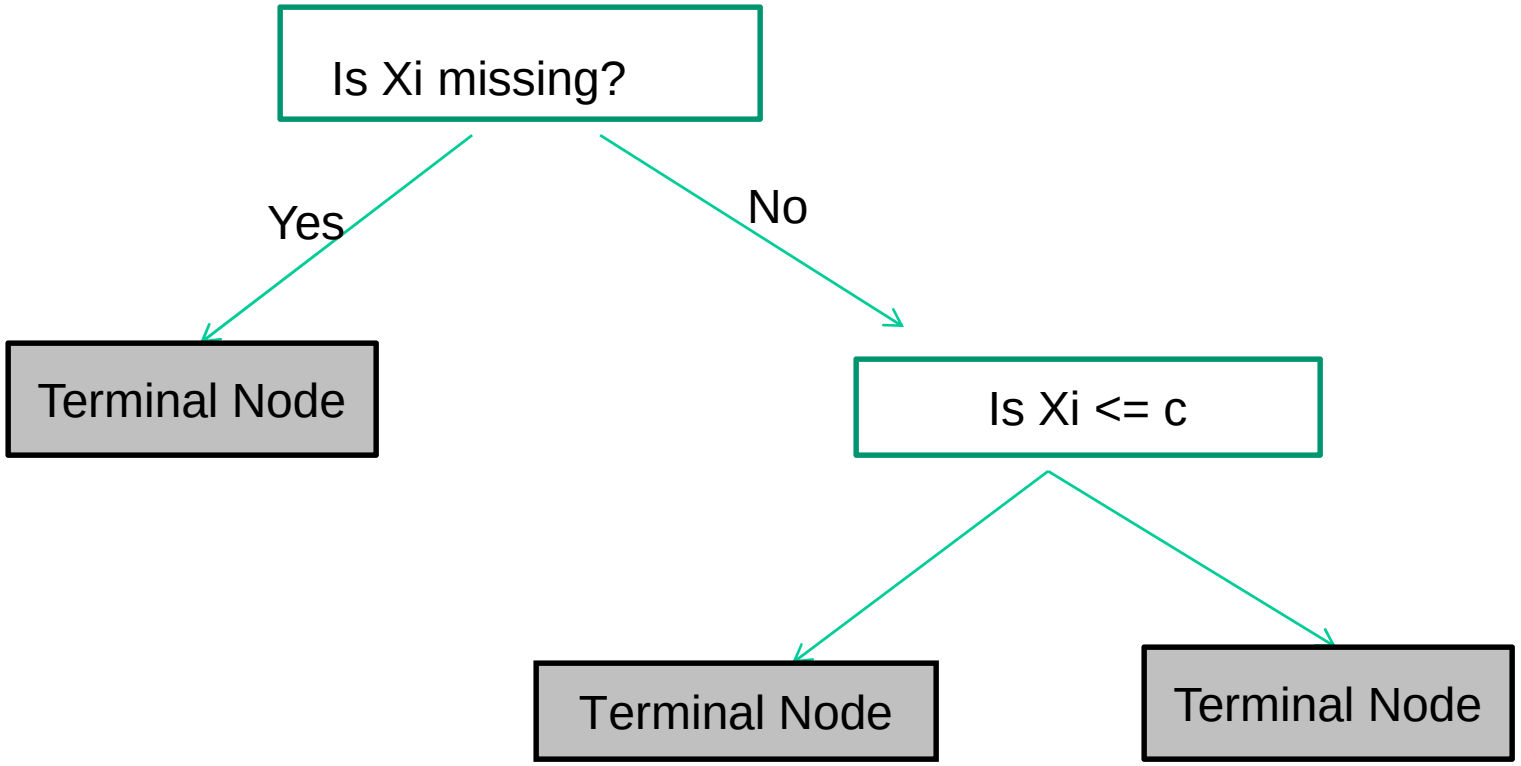
```
If  Xi == MISSING then go LEFT;
ELSE DO;
if CONDITION then go LEFT;
ELSE go RIGHT; END;
```

- The problem is that TreeNet permits the condition in the ELSE clause to involve a variable other than Xi

2-node trees in TreeNet

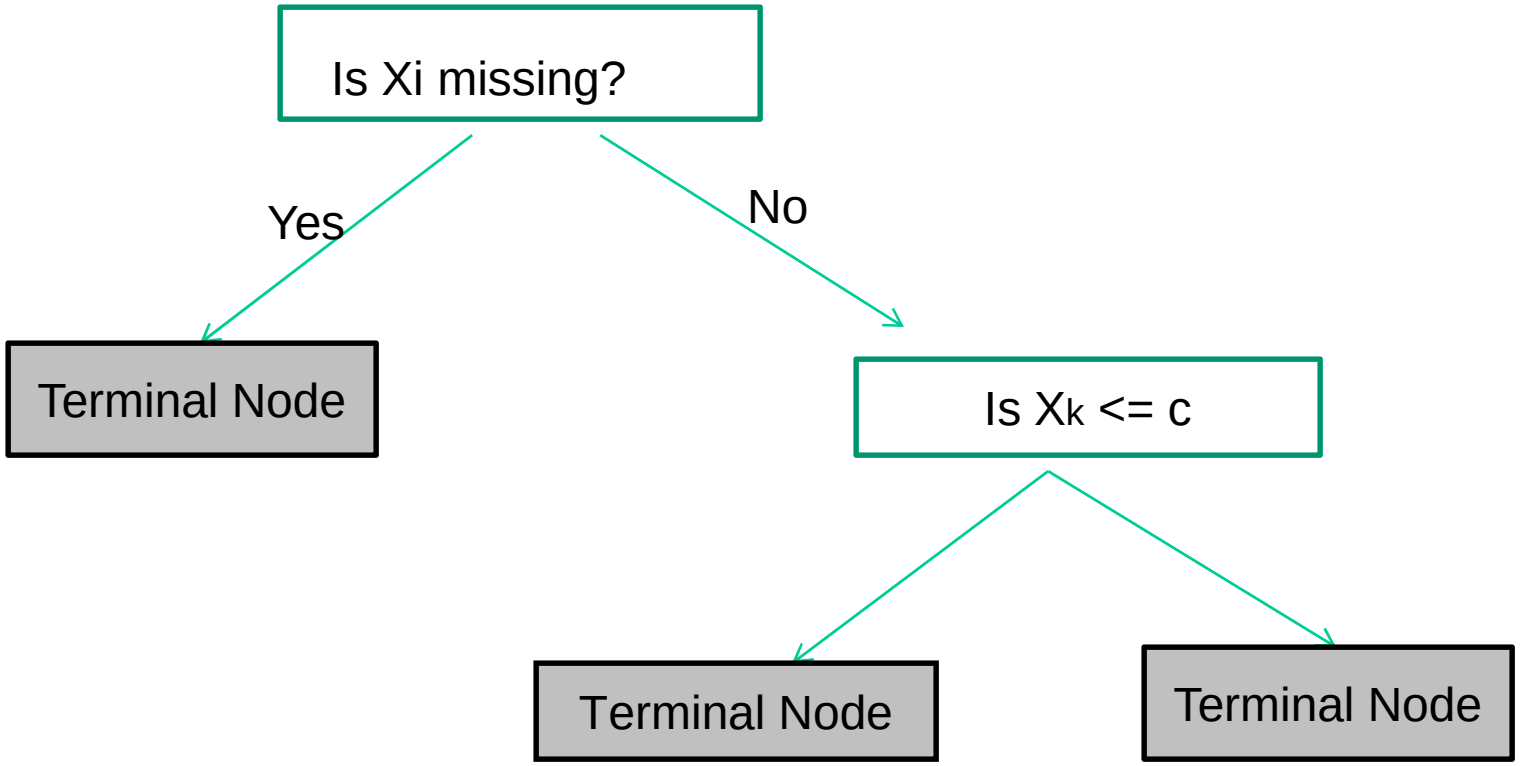
- Although the user may request 2-node trees in a TreeNet model if missing values are present for all predictors then the smallest possible tree that can be grown contains at least 3-nodes.
- An example of such a split follows:

“Two-Node” Tree with Missings One variable in tree



Both internal nodes are split using the same variable X_i

“Two-Node” Tree with Missings
Two variables in tree

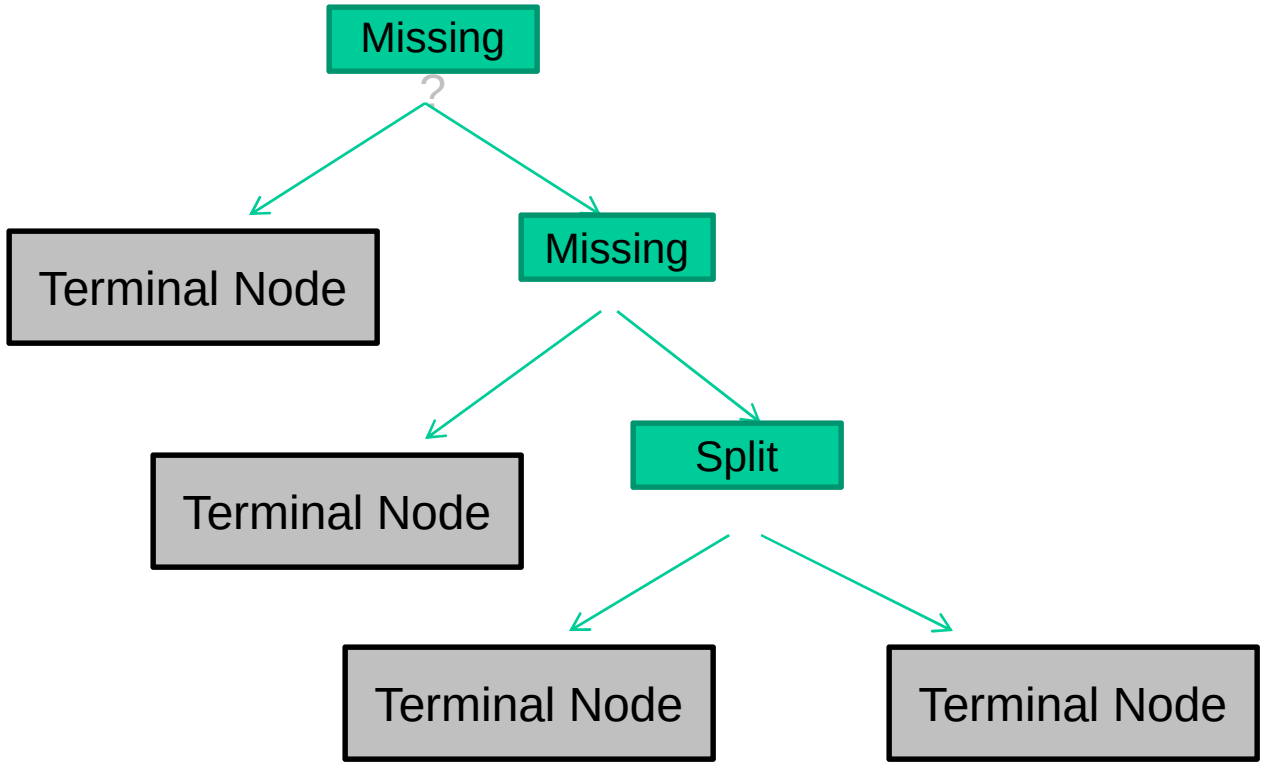


X_i (root node split) and X_k (internal split) are the two variables in the tree

2 node tree details

- A 2-node tree in TreeNet can never contain more than one split on a standard predictor. Therefore if the predictor is never missing the tree using this variable will have only two nodes
- However, the TreeNet mechanism does not “count” a split on a *missing value indicator* as a genuine split
- As a result “2-node” tree may contain any number of missing value indicator splits. The following tree is technically a “2-node” tree by TreeNet standards.

Allowable form for a TN “2-node” tree



2-node trees and Interactions

- The important point in this discussion is that in standard TreeNet we cannot guarantee the complete absence of interactions with 2-node trees
- The 2-node tree will allow interactions of any degree between missing value indicators and also a general interaction between a single predictor and missing value indicators
- To enforce literal non-interactivity we must rely on the ICL mechanism. If we specify that Xi is to enter the TreeNet as ADDITIVE then if we split the root using the MVI (missing value indicator) for Xi any subsequent split can use only Xi