Statistical Discovery.™ From SAS.

Form 712A Non-Government Disclosure (Contractor, Academic, UARC etc.)

1. Check the Event Website for the form deadline at www.mors.org.
2. If using the MORS Presenter Center all forms must be uploaded in the Presenter Center.
3. For other events e-mail the completed form to liz.marriott@mors.org.
4. If the presentation classification and or Disclosure Statement below is different than those entered in the MORS Presenter Center, correct the Presenter Center entries to ensure proper labeling of the Presentation.

PART I

Author Request - The following author(s) request authority to disclose the following presentation at the MORS Event below with subsequent publication in the MORS Event Report and posting on the MORS website if applicable.

Principal Author: Thomas A. Donnelly

Other Author(s):

Principal Author's Organization and complete mailing address:
SAS Institute Inc.
27 Farmingdale Ln
Newark, DE 19711

Principle Author Phone: 302-489-9291

Principle Author FAX:

Principle Author E-mail: tom.donnelly@jmp.com

Principal Author's Signature X: Thomas A. Donn

Principal Author's Organization and complete mailing address_Row_1

Date: 3 June 2021

MORS Event: 89th MORSS

Event Date(s): 21-24 June 2021

Presentation Type ☐ Plenary ☐ Course ☒ Tutorial ☐ Special Session ☐ Poster ☐ Demonstration

☐ Working/Composite/Distributed or Focus Group List All ☐ Other

Title of Presentation: Machine Learning – Using Robust Data Mining Methods

Presentation ID (if assigned): 56989

Classification ☐ SECRET ☐ SECRET//REL TO FVEY ☐ CONFIDENTIAL ☐ CONFIDENTIAL//REL TO FVEY

☒ UNCLASSIFIED ☐ UNCLASSIFIED W/FOUO ☐ Other

Distribution Statement ☒ A (Publicly Releasable) ☐ B ☐ C ☐ D ☐ E (see side 2 for definitions)

A.	This work was performed in connection with a government contract.	☐ YES (Complete Parts I, II & III)
B.	This presentation is based on material developed by the author as part of company/organization approved research e.g. IR&D and was NOT done under a government contract.	☐ YES (Complete Parts I, II & III)
C.	This presentation was NOT done under a government contract, contains no government information, is my own work and is approved for public release.	☒ YES (Complete Part I only)

INTRODUCTION TO MACHINE LEARNING USING ROBUST DATA MINING METHODS

89th MORSS
Webcast Tutorial 56989
June 21, 2021

Tom Donnelly, PhD, CAP
JMP Defense & Aerospace Team
Principal Systems Engineer
tom.donnelly@jmp.com
302-489-9291

TODAY

- Session 1: Machine Learning Intro
  - Honest Assessment Approach to prevent overfitting
- Session 2: Using Standard JMP
  - Regression, Partition, Neural & Predictor Screening
  - Creating variable subsets for validation, K-fold, Excluded row holdback, Other criterion
- Session 3: Using JMP Pro
  - Generalized Regression, BF & BT, Dual-layer NN, Boosted NN,
  - Validation column creation options
  - Model Comparison
  - Model Publishing – Formula Depot

QUESTIONS FROM END OF SESSION 1 IN AUGUST

1. Does Bootstrap Forest detect Interactions?
2. Difference between stratification and grouping?

Feel free to ask questions as we go along.

OUTLINE

- Resources
- Machine Learning from a Process Perspective
- Moving from Data to Understanding
- Model Overelaboration
- Honest Assessment to Prevent Overfitting
- Helicopter Surveillance – Supervised Learning Example
- Robust Machine Learning Strategy
- Countering Transnational Threats – Unsupervised Learning Example
- Apply Machine Learning to new types of data – Text & Data Streams
- Takeaways

RESOURCES

My Recorded Tutorials & Slide Decks at www.jmp.com/fedgov

These 9 videos cover predicitve analytics (including text exploration), data visualization, and "What's New in JMP 14?" topics.

Building Better Models Overview and Use of Honest Assessment	Neural Networks - Single Layer, Dual Layer, Boosted	All Graphs are Wrong - Some are Useful - Or view Xan Gregg's Original 2015 Discovery Summit Presentation
Regression Linear, Stepwise, Logistic, & All Possible	Generalized Regression Near Machine Learning Accuracy – More Explainable Model	What's New in JMP 14? JMP Learning Resources
Decision Trees Simple Partition, Bootstrap Forest, & Boosted Tree	Text Exploration Analyze Unstructured Free Text	Functional Data Explorer Modeling a "Stream" of Data – New in JMP 14

RECENT CHANGE TO COMMUNITY.JMP.COM

PUT THE LEARN JMP CONTENT ALL TOGETHER IN ONE PLACE

Discussions
Solve problems and share tips & tricks with other JMP users.

File Exchange
Download and share JMP add-ins, scripts, and sample data.

JMP Blogs
Read about a broad range of data analysis topics and posts that inform your JMP use.

Learn JMP
Extend your JMP skills with on-demand videos and JMP files.

JMP Wish List
We want to hear your ideas for improving JMP. Share them here.

JSL Cookbook
Building blocks of JSL code to reduce your coding workload.

JMP Users Groups
Meet up and discuss with other JMP users near you.

Discovery Summit
Info on upcoming Summits and materials from past events.

Community Help
Help with getting started, finding things, and how the Community works.

Find many topics not covered on the FedGov Users Resources Page
www.jmp.com/fedgov

Getting Started
Start here to learn the basic operation of JMP (recommended for all new users). Includes Welcome Kit and DOE Welcome Kit.

Short Videos
Short videos and guides to help you learn JMP. Includes STIPS (Statistical Thinking for Industrial Problem Solving) modules.

Tutorials
In-depth tutorial videos to help you learn analytical procedures. Includes Mastering JMP on-demand videos and materials.

Learning Paths
Curated learning paths to help you expand your knowledge of analytical topics, whether you're an advanced user, or just starting out.

Activities
Keep your skills sharp with these hands-on data challenges and activities.

Statistical Thinking for Industrial Problem Solving

A free online course

In virtually every field, deriving insights from data is central to problem solving, innovation and growth. But without an understanding of which approaches to use, and how to interpret and communicate results, the best opportunities will remain undiscovered.

That's why we created Statistical Thinking for Industrial Problem Solving. This online course is available – for free – to anyone interested in building practical skills in using data to solve problems better.

Have two minutes? Learn more.

Enroll now

https://www.[illegible].com/en_us/statistical-thinking.html

Statistical Thinking and Problem Solving

Statistical thinking is about understanding, controlling and reducing process variation. Learn about process maps, problem-solving tools for defining and scoping your project, and understanding the data you need to solve your problem.

Exploratory Data Analysis

Learn the basics of how to describe data with graphics and statistical summaries. Then, learn how to use interactive visualizations to communicate the story in your data. You'll also learn some core steps in preparing your data for analysis.

Quality Methods

Learn about tools for quantifying, controlling and reducing variation in your product, service or process. Topics include control charts, process capability and measurement systems analysis.

Decision Making With Data

Learn about tools used for drawing inferences from data. In this module you learn about statistical intervals and hypothesis tests. You also learn how to calculate sample size and see the relationship between sample size and power.

Correlation and Regression

Learn how to use scatterplots and correlation to study the linear association between pairs of variables. Then, learn how to fit, evaluate and interpret linear and logistic regression models.

Design of Experiments

In this introduction to statistically designed experiments (DOE), you learn the language of DOE, and see how to design, conduct and analyze an experiment in JMP.

Predictive Modeling and Text Mining

Learn how to identify possible relationships, build predictive models and derive value from free-form text.

RESOURCES

- ***Demystifying Data Science*** presented at DATAWorks 2018

  by Prof. Alyson Wilson, NC State Laboratory for Analytical Sciences

https://dataworks2018.testscience.org/wp-content/uploads/sites/8/2018/03/demystifying-data-science_Alyson-Wilson.pdf

Data science is the new buzz word – it is being touted as the solution for everything from curing cancer to self-driving cars. How is data science related to traditional statistics methods? Is data science just another name for “big data”? In this mini-tutorial, we will begin by discussing what data science is (and is not). We will then discuss some of the key principles of data science practice and conclude by examining the classes of problems and methods that are included in data science.

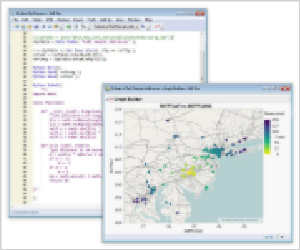
Dr. Laura Freeman, Virginia Tech, Director Intelligent Systems Lab, Hume Center for National Security and Technology – August 13th

https://www.jmp.com/en_gb/whitepapers/jmp/jmp-synergies-using-jmp-and-jmp-pro-with-python-and-r.html

Using JMP and JMP Pro With Python and R

JMP Synergies With Open Source

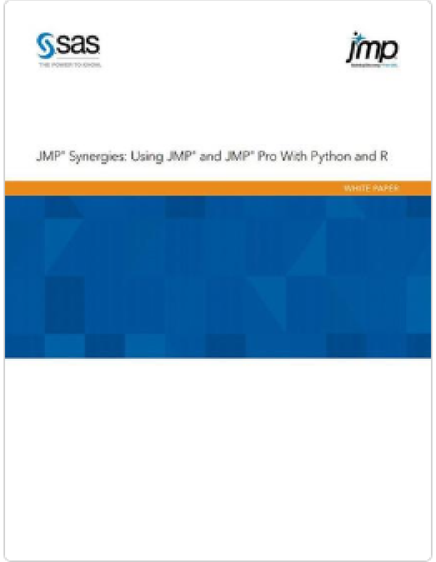
By Ruth Hummel, JMP



JMP is a standalone, full-featured data visualization and statistical analysis software from SAS for the Windows and Mac desktop. JMP has the interactivity and dynamic linkage that makes data exploration exciting, insightful and contains many advanced analytical options, fully satisfying the needs of data explorers and analysts. Still, there may be occasions where you'll want (or need) to use JMP in conjunction with open source tools, like Python or R.

In addition to providing you with the basics, this paper introduces the Python scoring-code generation, the Python in JMP scripting and the R in JMP scripting. You'll also discover sample code and advanced examples that will make using the connections and add-ins provided in JMP easy.

Whatever your motivation for connecting open source (or other) tools with JMP software's GUI, this guide will help you to get started using the Python and R connections in JMP.



Download Now

Download white paper

JMP® ANALYTIC WORKFLOW

ODBC

Compile Data

Prepare/Curate Data

Explore & Visualize Data

Analyze Data/Build Models

Organize Results/Findings

Share Results/Communicate

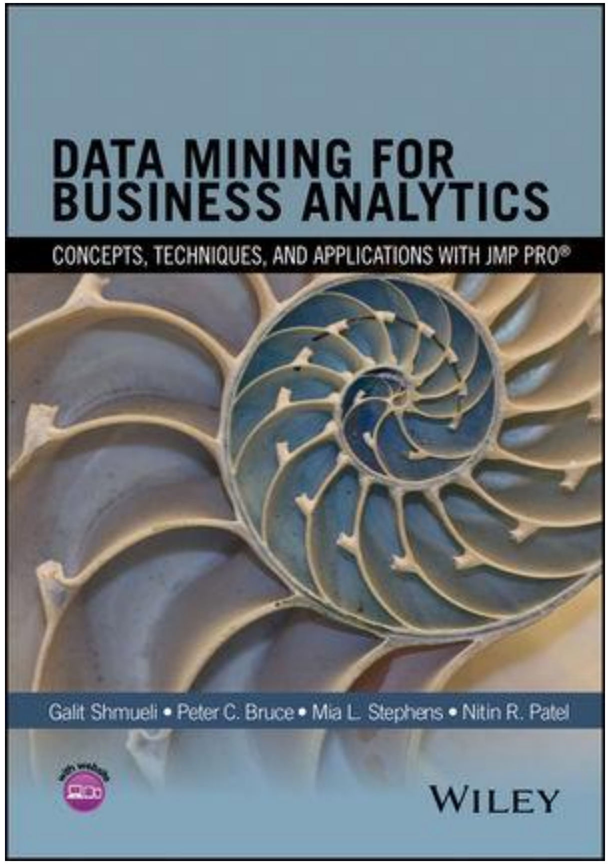
HTML

RESOURCES | GO-TO BOOK TO TEACH DATA SCIENCE IN R - (ALYSON WILSON)

G. Shmueli, P. Bruce, I. Yahav, N. Patel, K. Lichtendahl (2018). ***Data Mining for Business Analytics: Concepts, Techniques, and Applications in R***. John Wiley & Sons.

G. Shmueli, P. Bruce, P. Gedeck, N. Patel (2019). ***Data Mining for Business Analytics: Concepts, Techniques, and Applications in Python.*** John Wiley & Sons.

G. Shmueli, P. Bruce, M. Stephens, N. Patel (2017). ***Data Mining for Business Analytics: Concepts, Techniques, and Applications with JMP Pro®***. John Wiley & Sons.



MACHINE LEARNING FROM A PROCESS PERSPECTIVE

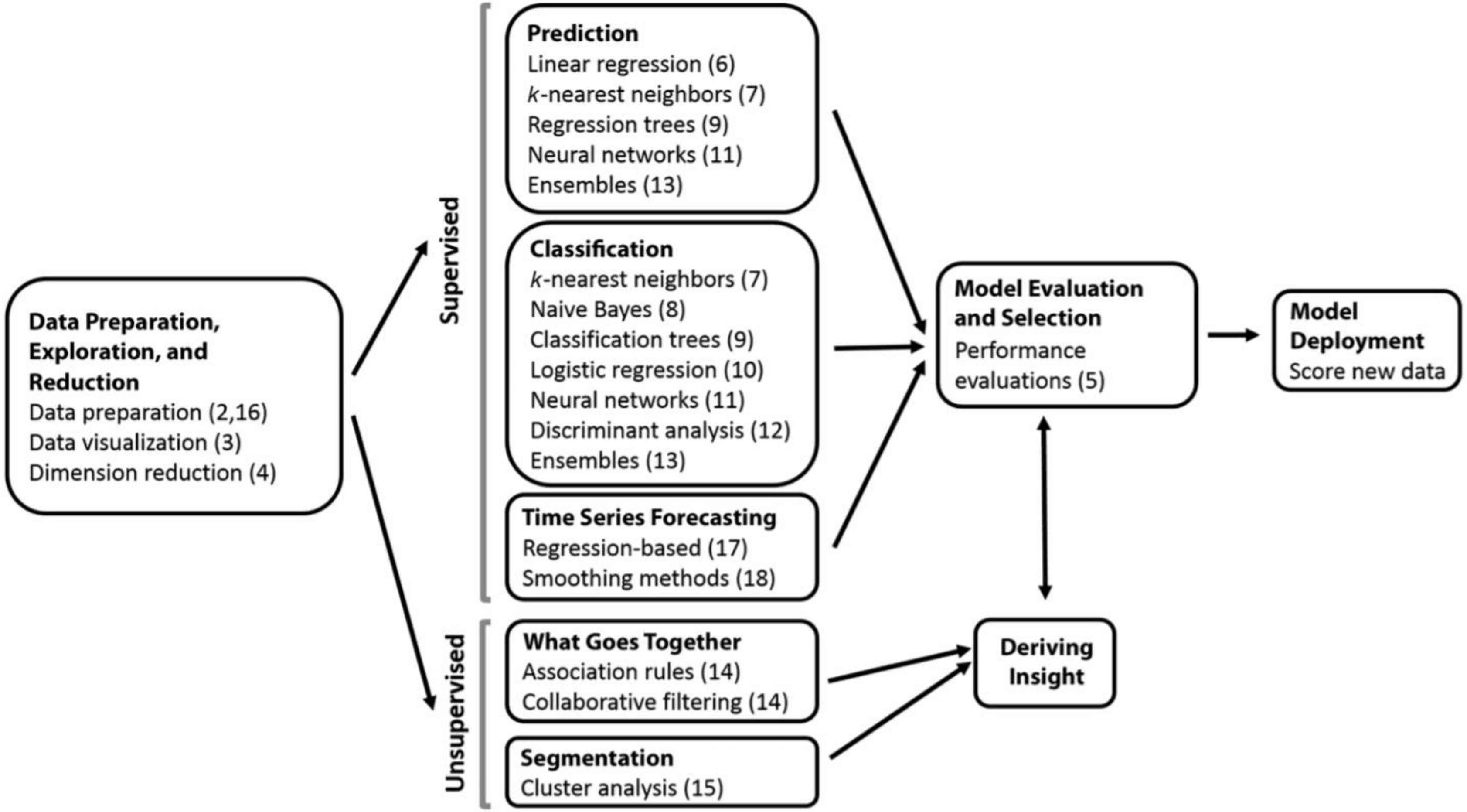


FIGURE 1.2 DATA MINING FROM A PROCESS PERSPECTIVE. NUMBERS IN PARENTHESES INDICATE CHAPTER NUMBERS

So, do we throw the book at our problem?

Maybe not the whole book, but perhaps the prediction and classification sections.

Prediction
Linear regression (6)
k-nearest neighbors (7)
Regression trees (9)
Neural networks (11)
Ensembles (13)

Classification
k-nearest neighbors (7)
Naive Bayes (8)
Classification trees (9)
Logistic regression (10)
Neural networks (11)
Discriminant analysis (12)
Ensembles (13)

Goal is to streamline workflow to rapidly identify the top contending modeling methods.

Rather than iteratively fitting all models, simultaneously fit just the desired ones and compare their performance.

JMP Pro 16 Adds *Model Screening*

Run selected models all at once, then view ranked performance

Method

- ☑ Decision Tree
- ☑ Bootstrap Forest
- ☑ Boosted Tree
- ☑ K Nearest Neighbors
- ☐ Naive Bayes
- ☑ Neural
- ☑ Support Vector Machines
- ☑ Discriminant
- ☑ Fit Least Squares
- ☑ Fit Stepwise
- ☑ Logistic Regression
- ☑ Generalized Regression
- ☐ Partial Least Squares
- ☐ XGBoost

Even run XGBoost via a JMP Addin

Options

- ☐ Remove Live Reports
- ☐ Log Methods
- Time Limit Each .
- Set Random Seed .

Folded Crossvalidation

Fit repeatedly with sequenced folds.

- ☐ K Fold Crossvalidation K 5
- ☐ Nested Crossvalidation K 4
- Repeated K Fold 0

Modeling Options

- ☐ Add Two Way Interactions
- ☐ Add Quadratics
- ☐ Informative Missing
- ☐ Additional Methods

Test

Method	Details	N	RSquare	RASE
Neural Boosted		998	0.8970	0.13810
Bootstrap Forest		998	0.8058	0.18962
Fit Stepwise	2FI Quad	998	0.7973	0.19373
Generalized Regression Lasso	2FI Quad	998	0.7470	0.21645
Boosted Tree		998	0.7193	0.22797
Decision Tree		998	0.6766	0.24473
Fit Least Squares	2FI Quad	998	0.6303	0.26164
Fit Stepwise		998	0.5946	0.27400
Fit Least Squares		998	0.4922	0.30666
Generalized Regression Lasso		998	0.4922	0.30666
K Nearest Neighbors		998	0.3808	0.33861
Support Vector Machines		998	0.2043	0.38385

Select Dominant Run Selected Save Script Selected

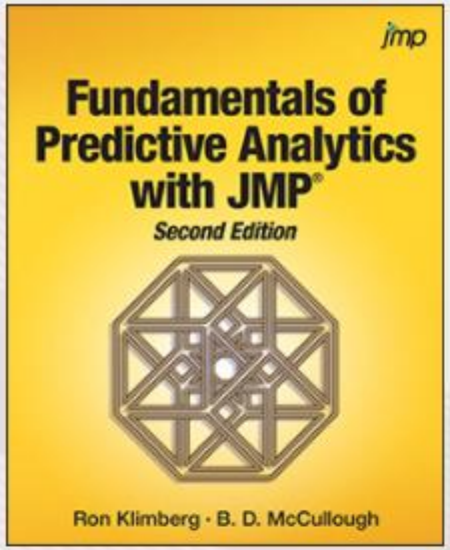
JMP Pro can Publish models in Python, C, SAS, SQL, JavaScript

jmp STATISTICAL DISCOVERY FROM SAS

RESOURCES

MY FAVORITE BOOK TO LEARN MACHINE LEARNING METHODS

Go to www.jmp.com/books for a 20% discount use code “SASCBP20”



Fundamentals of Predictive Analytics with JMP®, Second Edition

Ron Klimberg
B. D. McCullough

In Stock

Publisher: SAS Institute
Copyright Date: December 2016

- **Machine Learning**: focused on prediction, based on known properties learned from the training data.
- **Data Mining**: focused on the discovery of (previously) unknown properties in the data.
- Data Mining + Machine Learning are currently being rebranded as **Artificial Intelligence**.

“Why is a 4-star talking to a roomful of analysts?”

“I’ve got data.

Faster

What I need is information.
More than that I need knowledge.
And, more than that I need understanding.
So, I can take action.”

Admiral James “Sandy” Winnefeld Jr. (retired)
Vice Chairman of the Joint Chiefs of Staff (2011-2015)
Speaking at MORS MDA Workshop, Point Loma, CA, May 2011

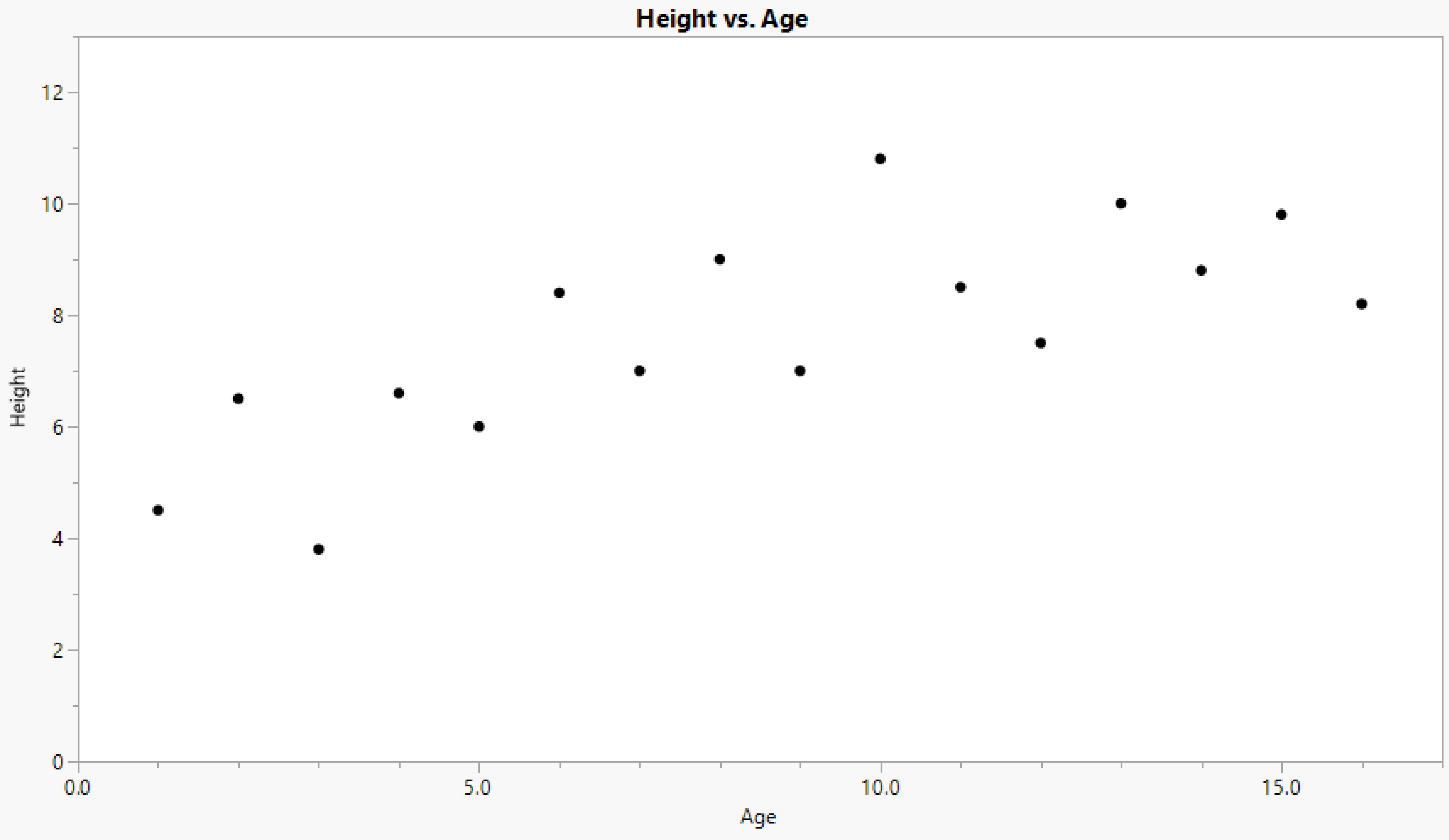
All models are wrong, but some are useful.

George E. P. Box (1979)

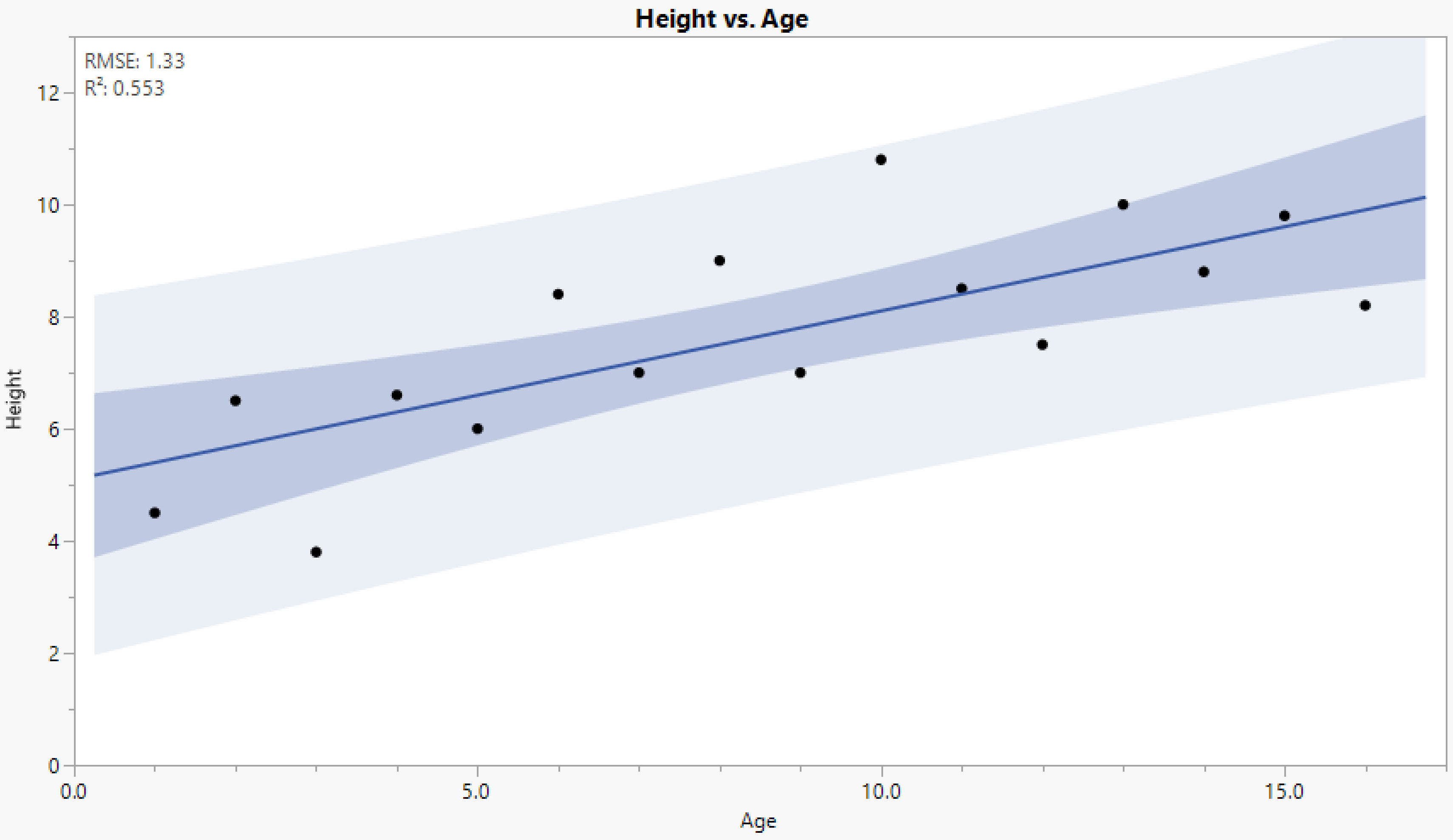
Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. … **overelaboration** and **overparameterization** is often the mark of mediocrity.

George E. P. Box (1976)

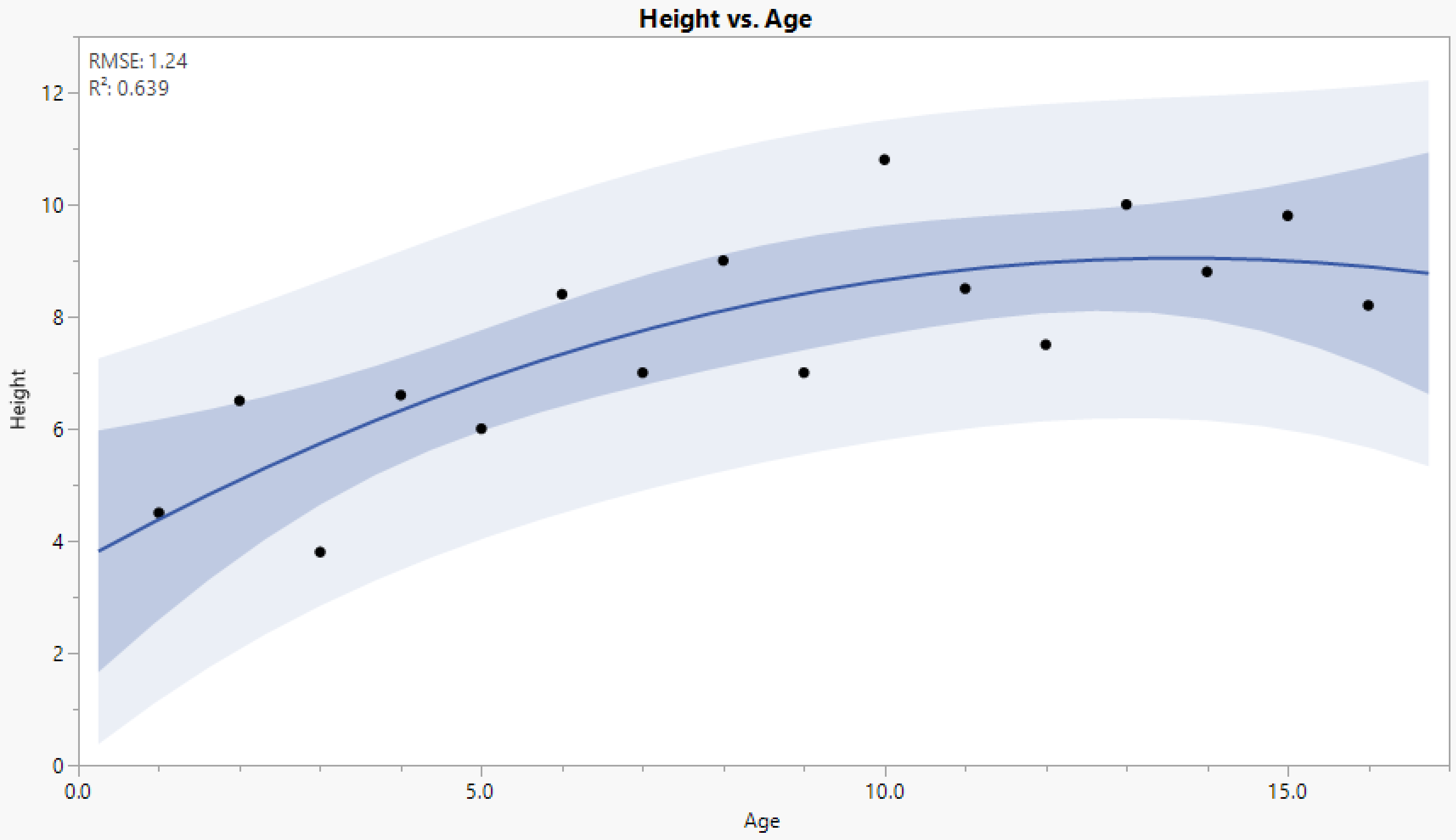
Overelaboration in Modeling



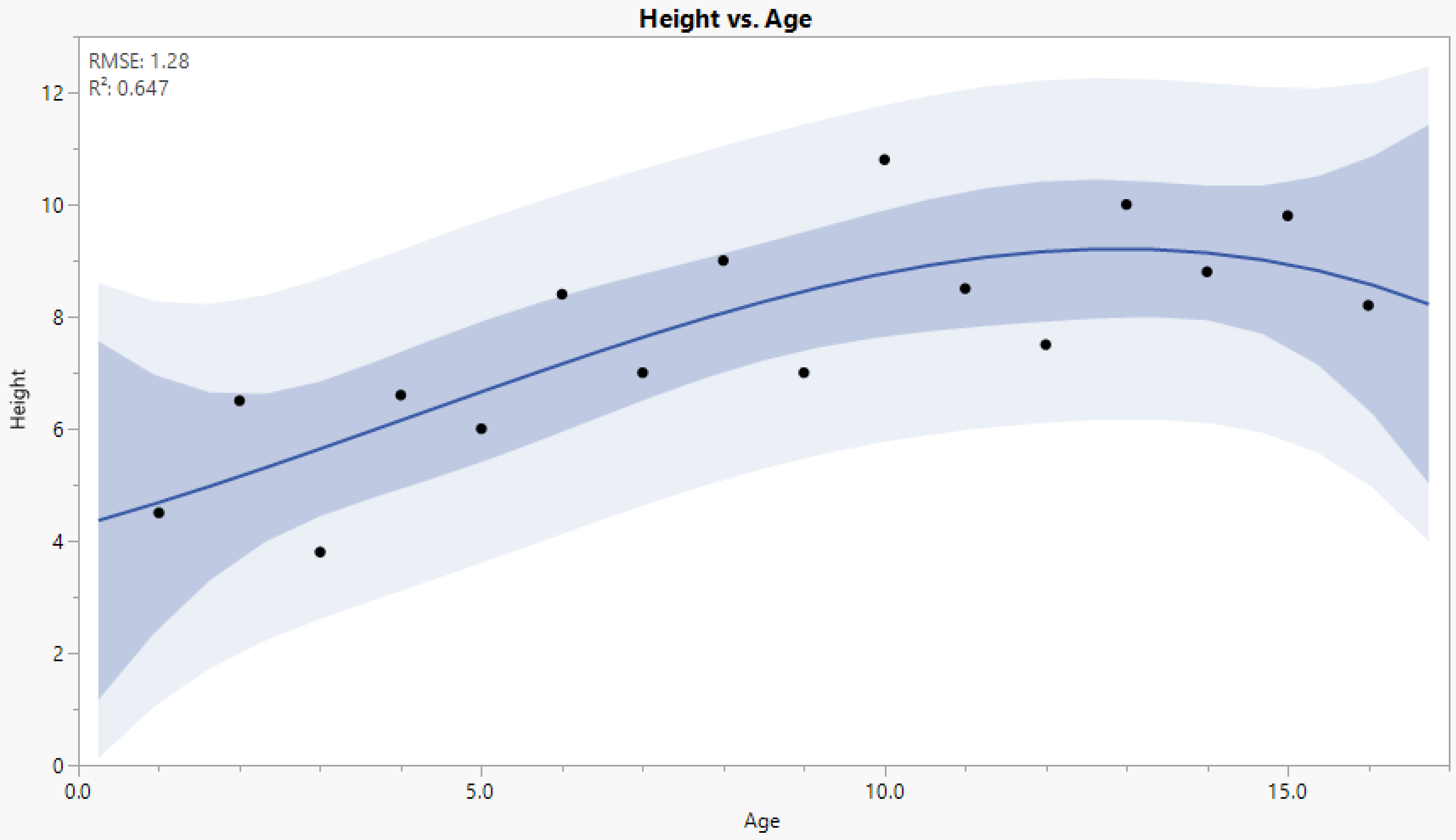
Overelaboration in Modeling



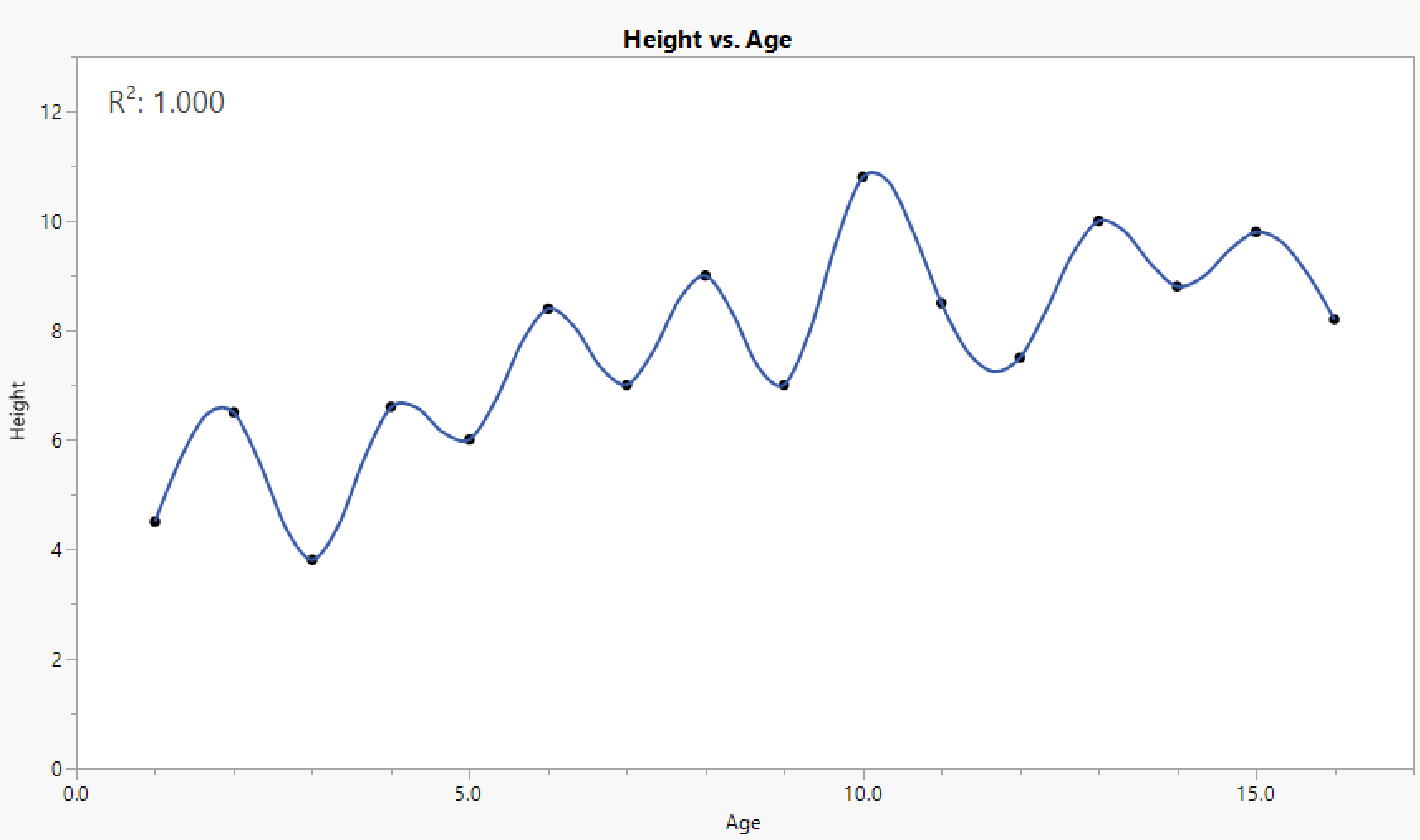
Overelaboration in Modeling

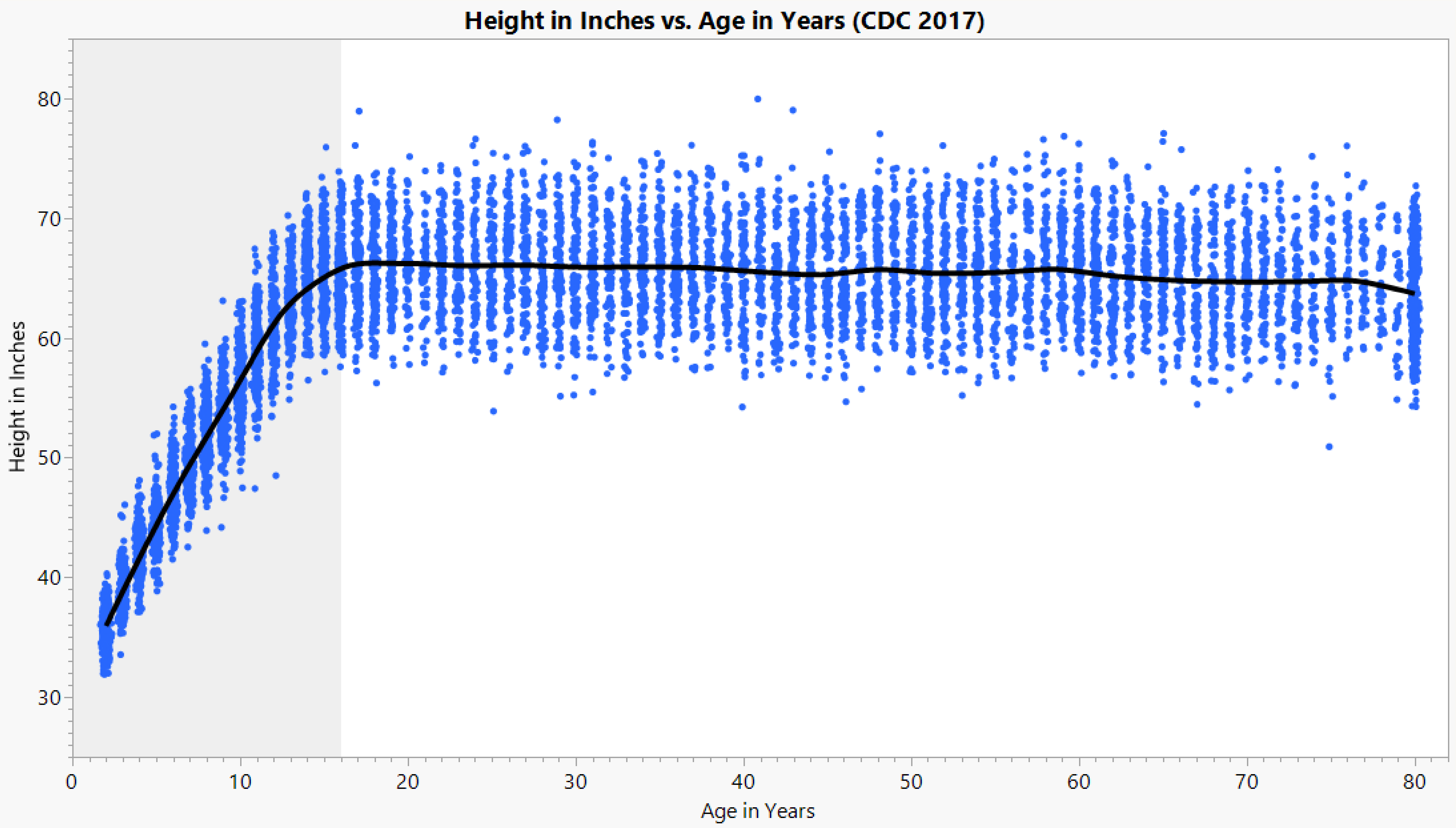


Overelaboration in Modeling



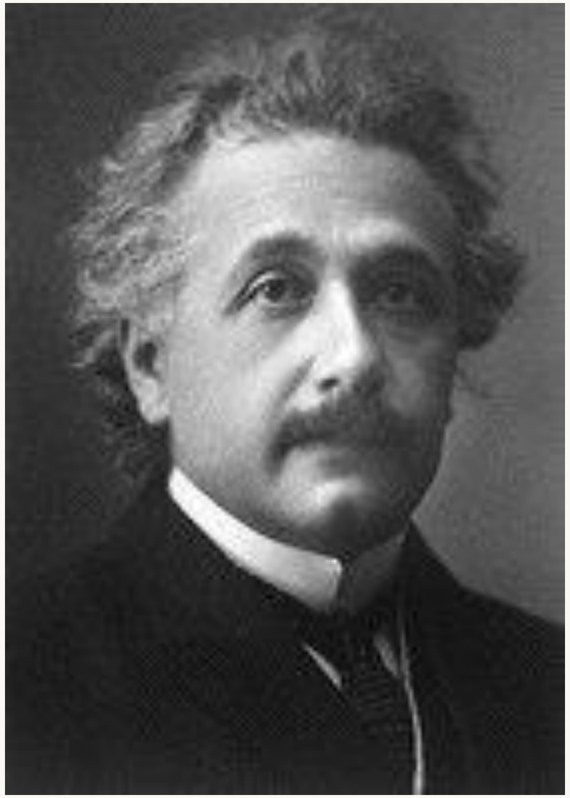
Overelaboration in Modeling

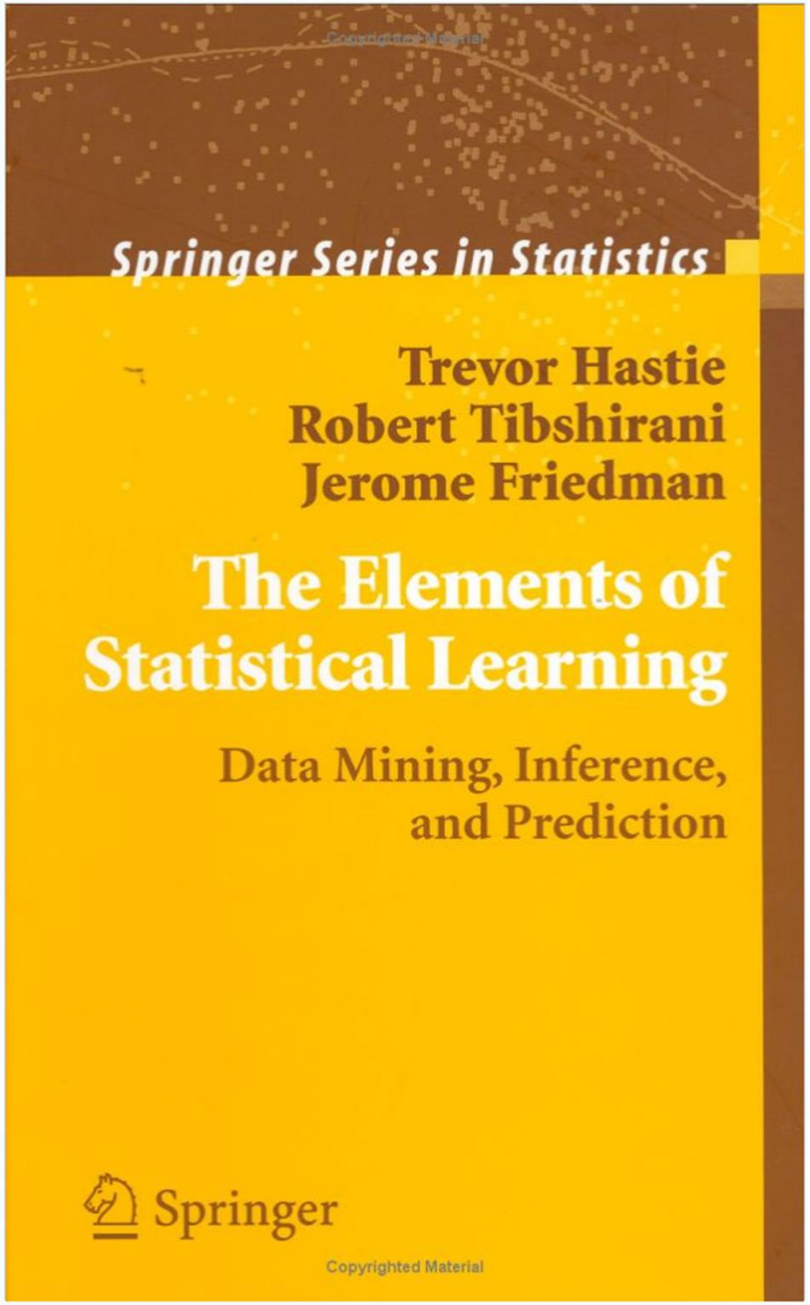



Height in Inches vs. Age in Years (CDC 2017)
Height in Inches
80
70
60
50
40
30
0
10
20
30
40
50
60
70
80
Age in Years

“Everything should be made as simple as possible, but not simpler.”

Attributed to Albert Einstein (1950)





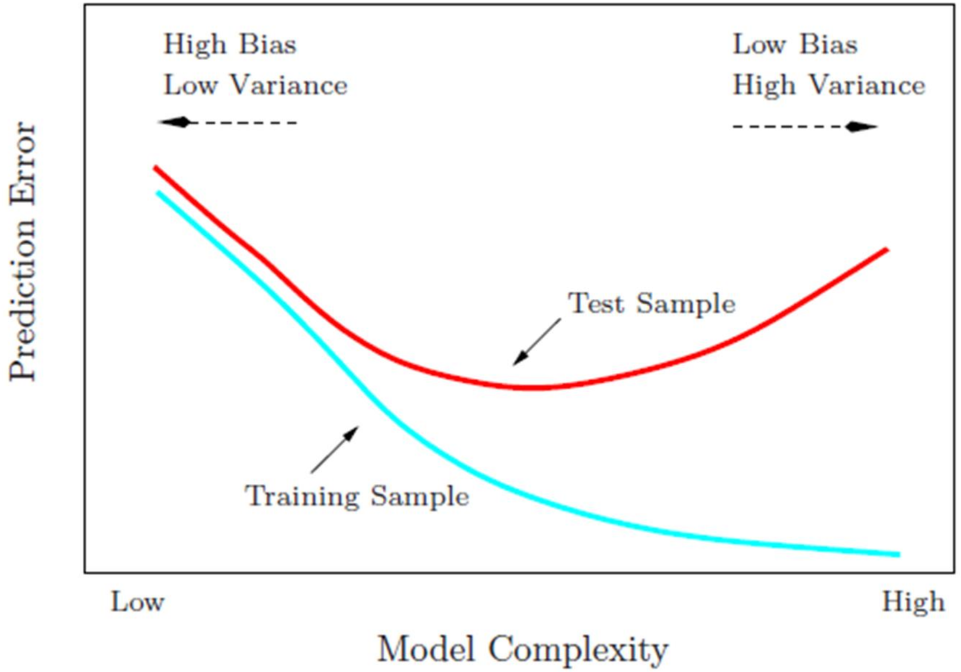


FIGURE 2.11. *Test and training error as a function of model complexity.*

It is difficult to give a general rule on how to choose the number of observations in each of the three parts, as this depends on the signal-to-noise ratio in the data and the training sample size. A typical split might be 50% for training, and 25% each for validation and testing:

The *bias error* is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).

The *variance* is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise. in the training data, rather than the intended outputs (overfitting).

“One can fit the data from a process but not necessarily fit the process from which the data come.” – Bob Wheeler

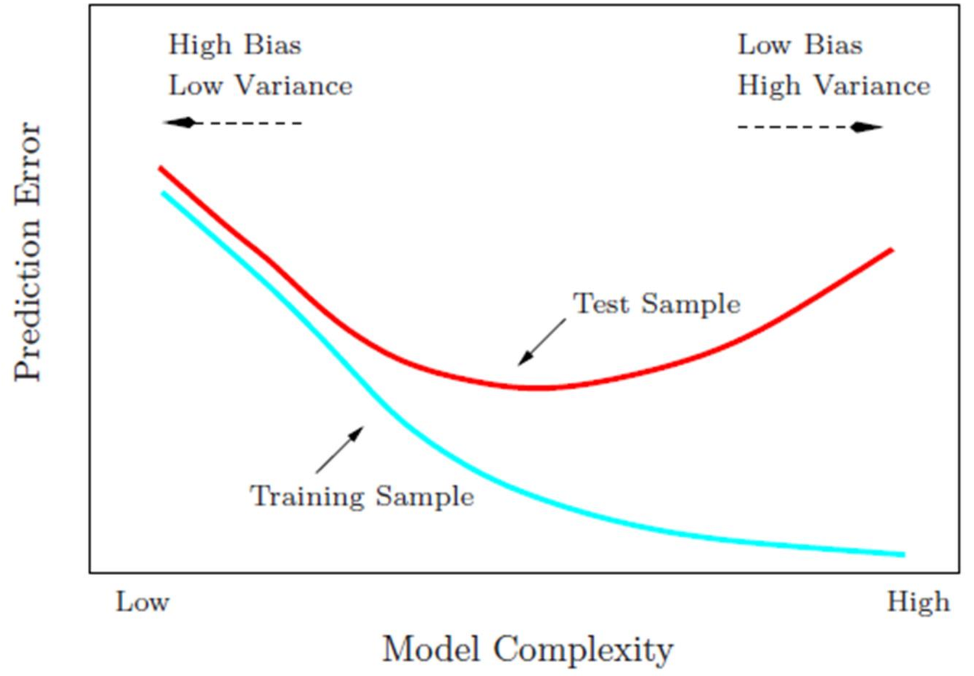
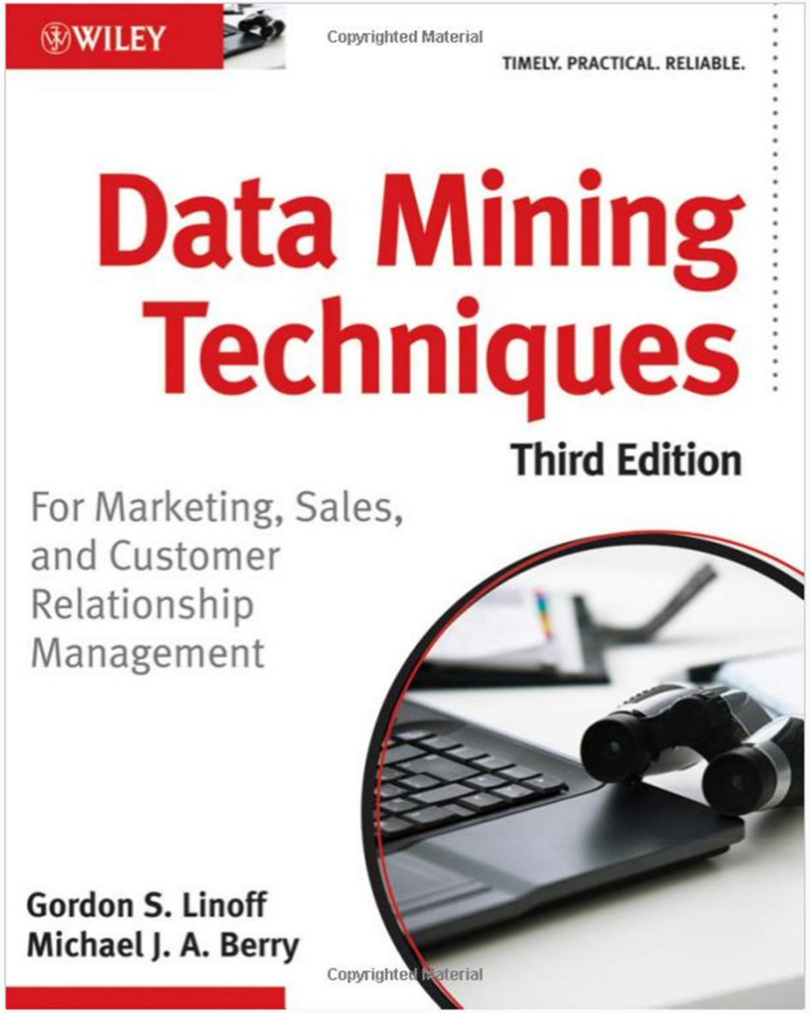


FIGURE 2.11. *Test and training error as a function of model complexity.*

It is difficult to give a general rule on how to choose the number of observations in each of the three parts, as this depends on the signal-to-noise ratio in the data and the training sample size. A typical split might be 50% for training, and 25% each for validation and testing:



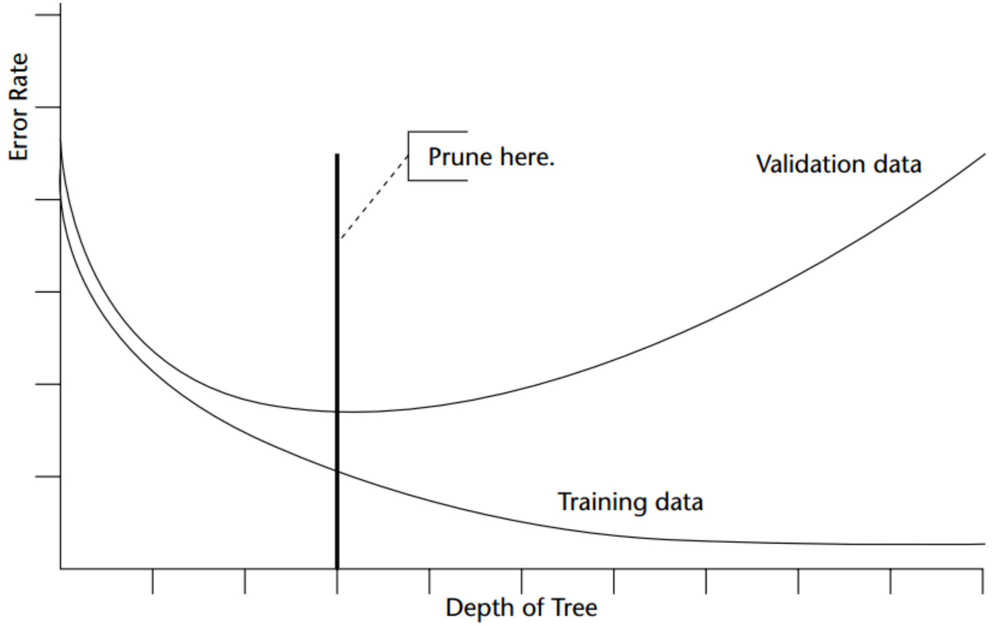


Figure 6.7 Pruning chooses the tree whose miscalculation rate is minimized on the validation set.

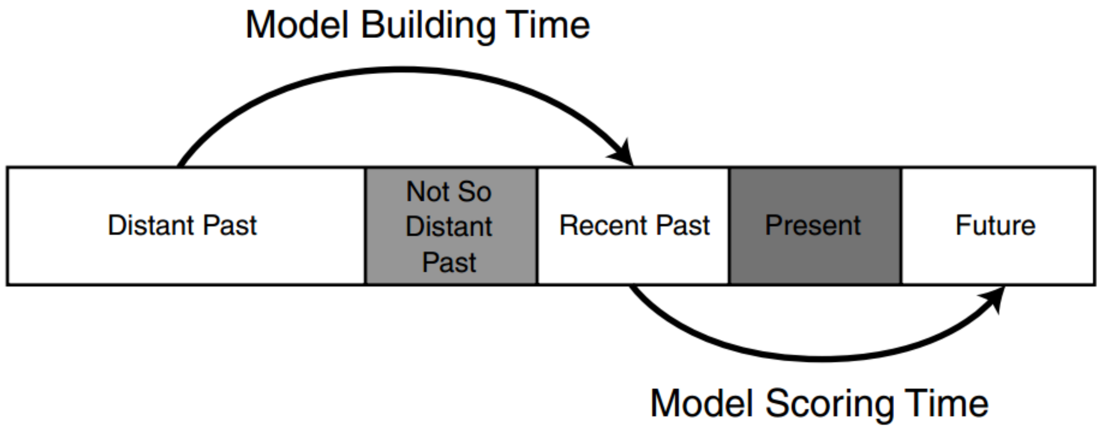
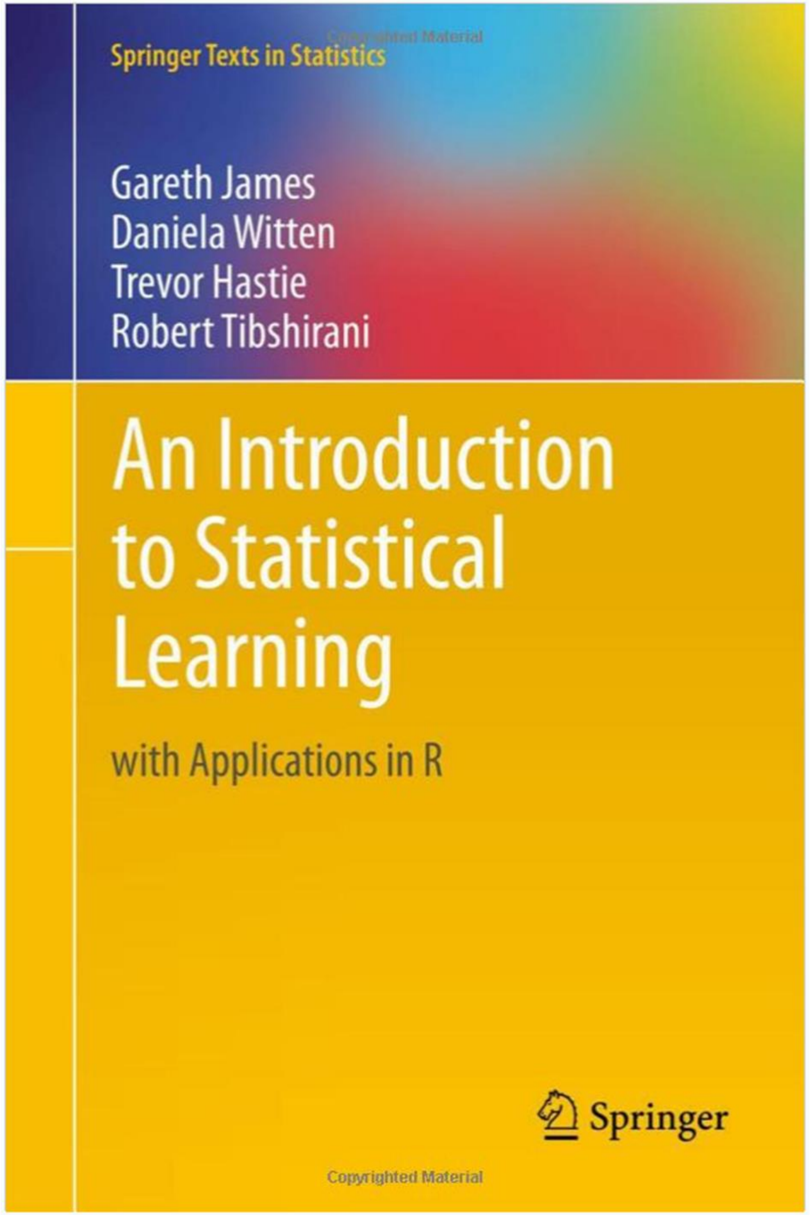


Figure 3.7 Data from the past mimics data from the past, present, and future.



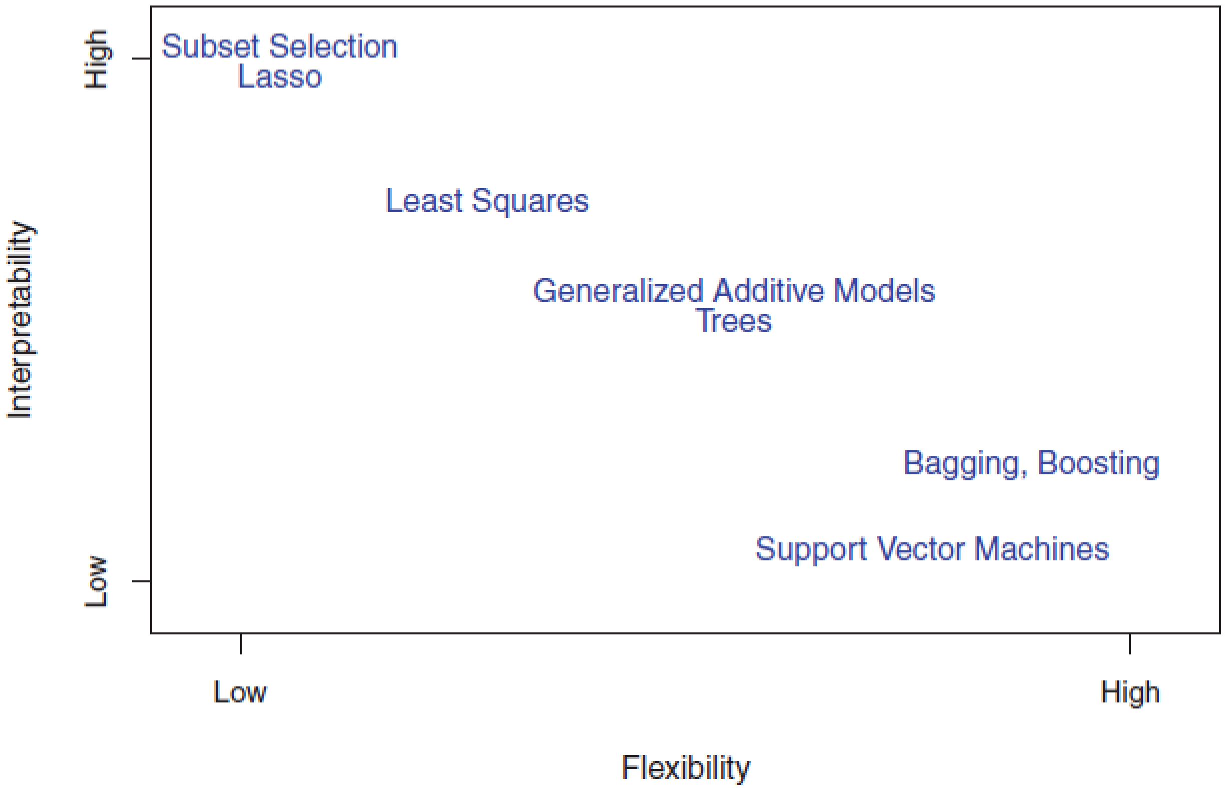
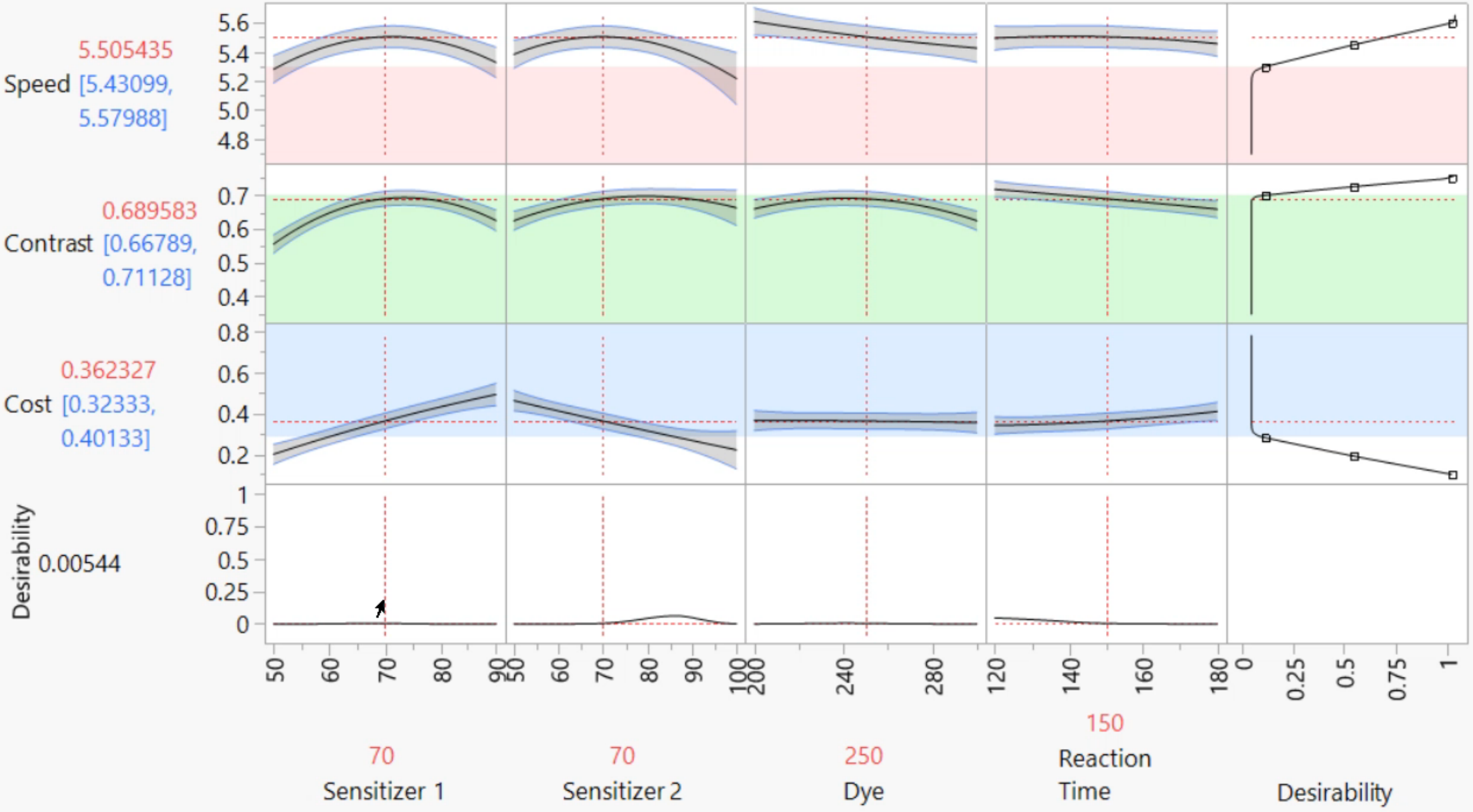


FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

USE JMP TRADE-OFF AND OPTIMIZATION

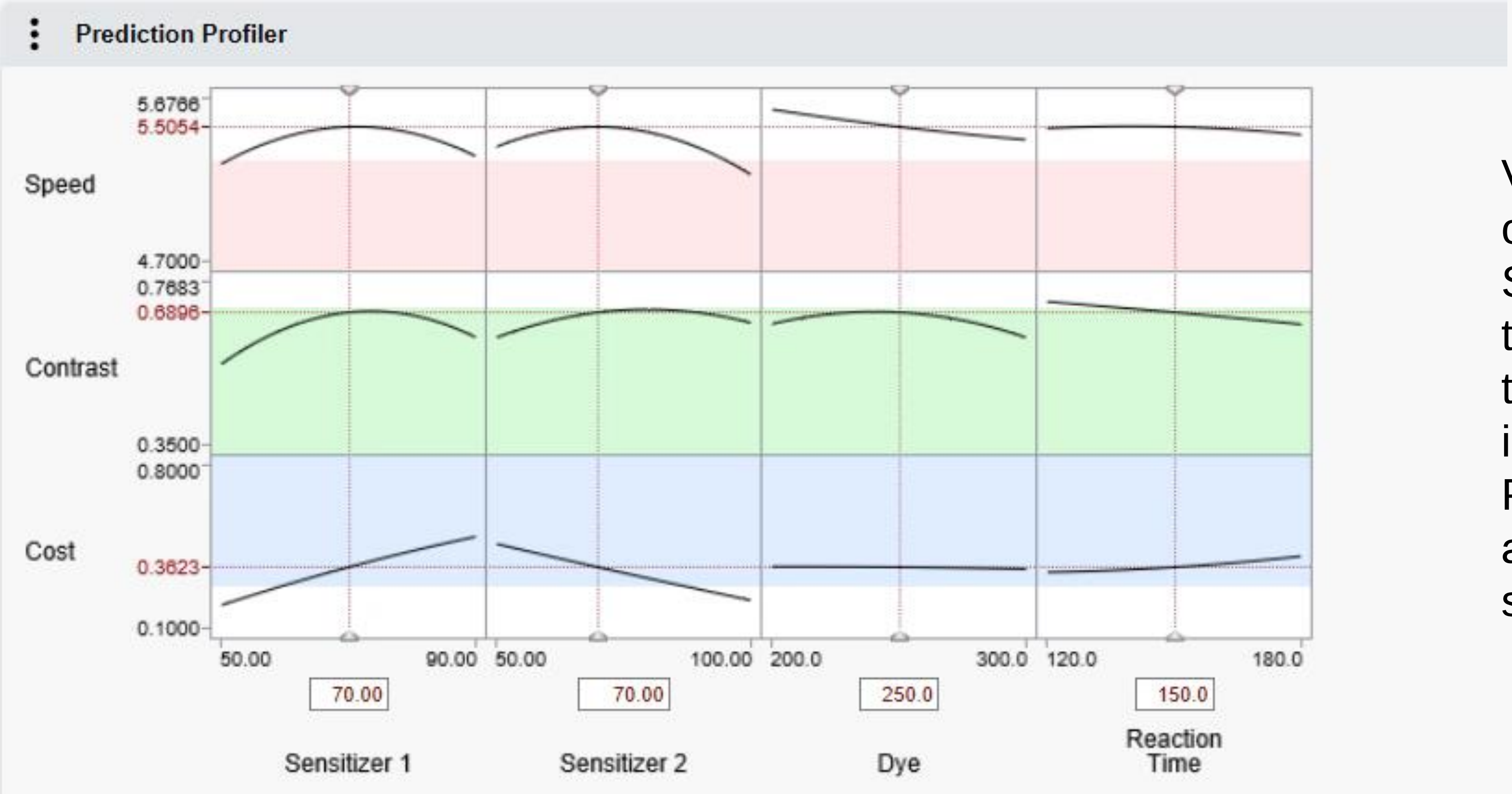
Prediction Profiler



Remembered Settings

Setting	Sensitizer 1	Sensitizer 2	Dye	Reaction Time	Speed	Contrast	Cost	Desirability
Equal Importance Opt	80.753574	91.269729	250.57625	120	5.3542877	0.7466933	0.2504014	0.347702
Mid Point Settings	70	70	250	150	5.5054353	0.6895831	0.3623274	0.004875
Cost 6X Speed & Contrast	84.016038	93.725925	283.02514	120	5.2902084	0.72549	0.1991539	0.214425
Opt Spd3X-Cntr1X-Cost6X	81.958309	90.706277	286.82246	120	5.3269582	0.7177857	0.2211116	0.264298

SHARE RESULTS ON JMP PUBLIC OR JMP LIVE



Remembered Settings

	Setting	Sensitizer 1	Sensitizer 2	Dye	Reaction Time	Speed	Contrast	Cost
Apply	Equal Importance Opt	80.753574	91.269729	250.57625	120	5.3542877	0.7466933	0.2504014
Apply	Mid Point Settings	70	70	250	150	5.5054353	0.6895831	0.3623274
Apply	Cost 6X Speed & Contrast	84.016038	93.725925	283.02514	120	5.2902084	0.72549	0.1991539
Apply	Opt Spd3X-Cntr1X-Cost6X	81.958309	90.706277	286.82246	120	5.3269582	0.7177857	0.2211116

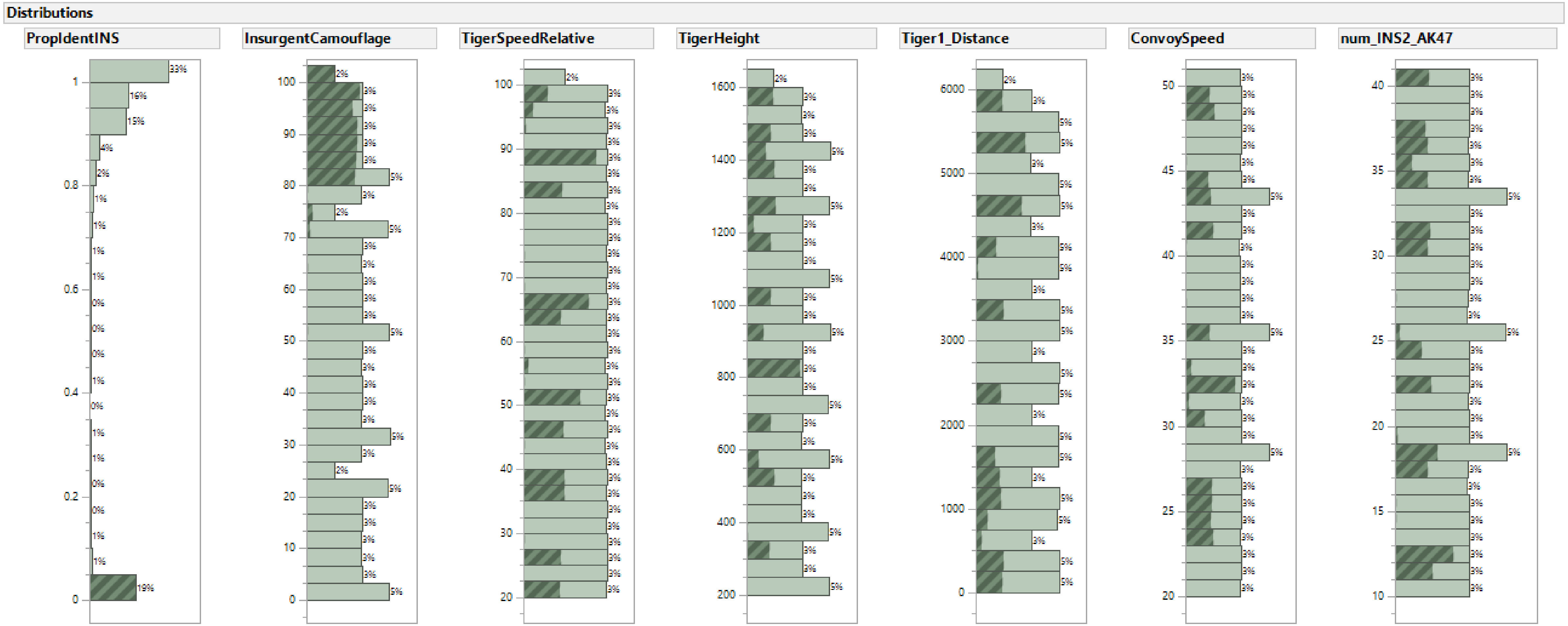
View optimizations on your phone. Scan the QR code to launch browser, then use finger to interact with the Prediction Profiler and to “Apply” saved settings.

SURROGATE MODELING OF A COMPUTER SIMULATION
HELICOPTER SURVEILLANCE – IDENTIFYING INSURGENTS

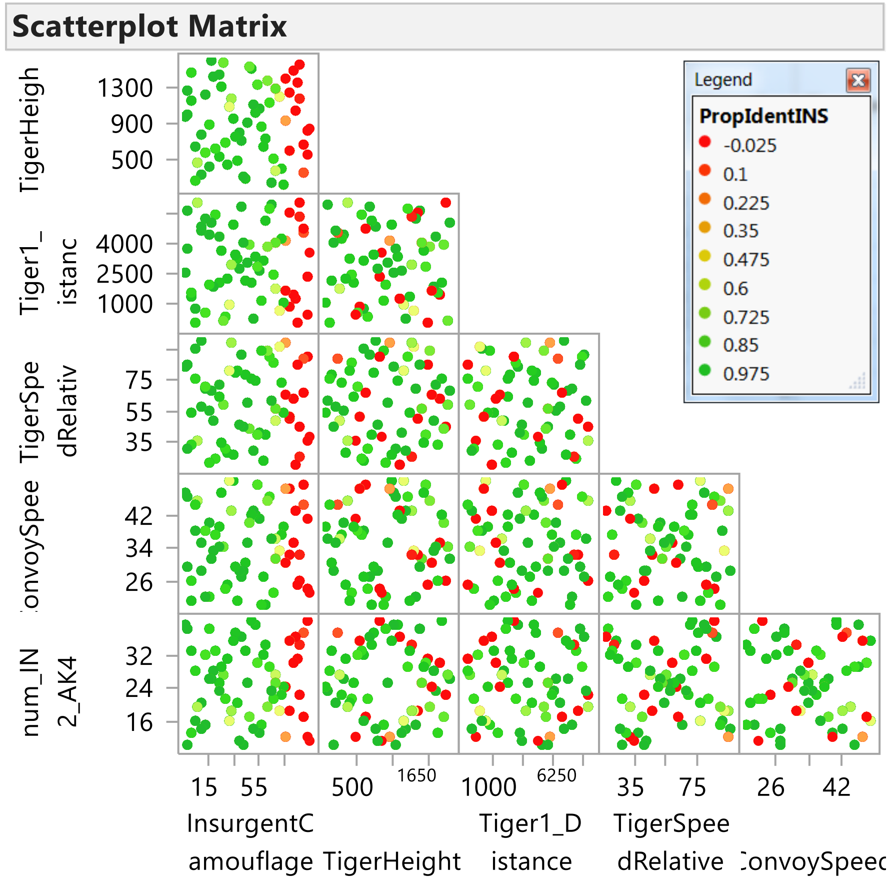
- 2009 International Data Farming Workshop - IDFW21, Lisbon, Portugal
- Largely German team (6 of 8) – their simulation
- 6500 simulations run overnight on cluster in Frankfurt
  - Space Filling Design of Experiments (DOE)
  - 65 unique combinations of 6 factors (each factor at 65 levels)
  - each case had 96 to 100 replications (lost a few)
- Response = Proportion of Insurgents Identified =

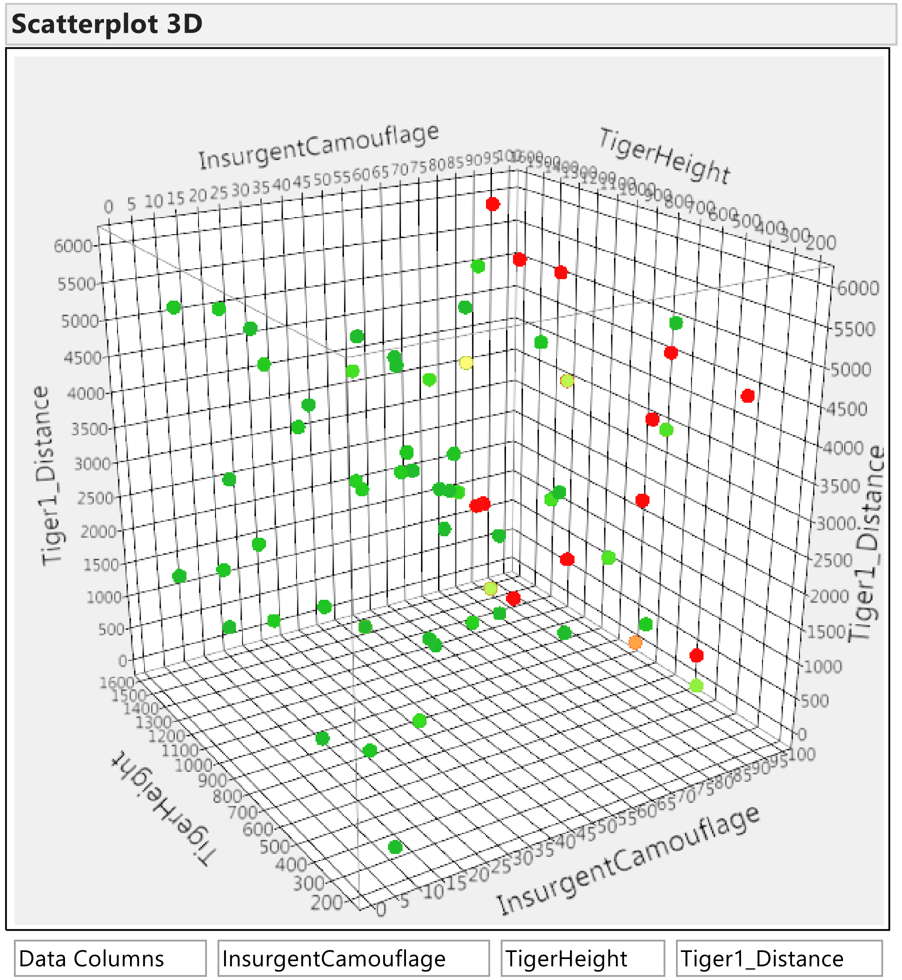
  PropIdentINS Data bounded between 0 and 1
- Explore data visually first
- Fit many different models using “Train, Validate (Tune), Test” subsets
- Compare Actual vs. Predicted for Test Subsets

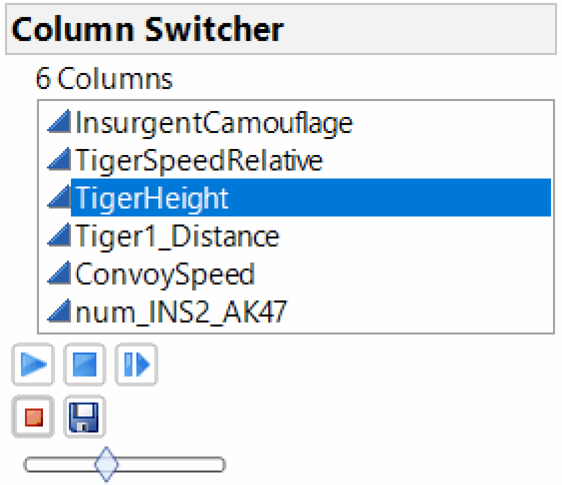
DISTRIBUTIONS OF 1 RESPONSE AND 6 FACTORS



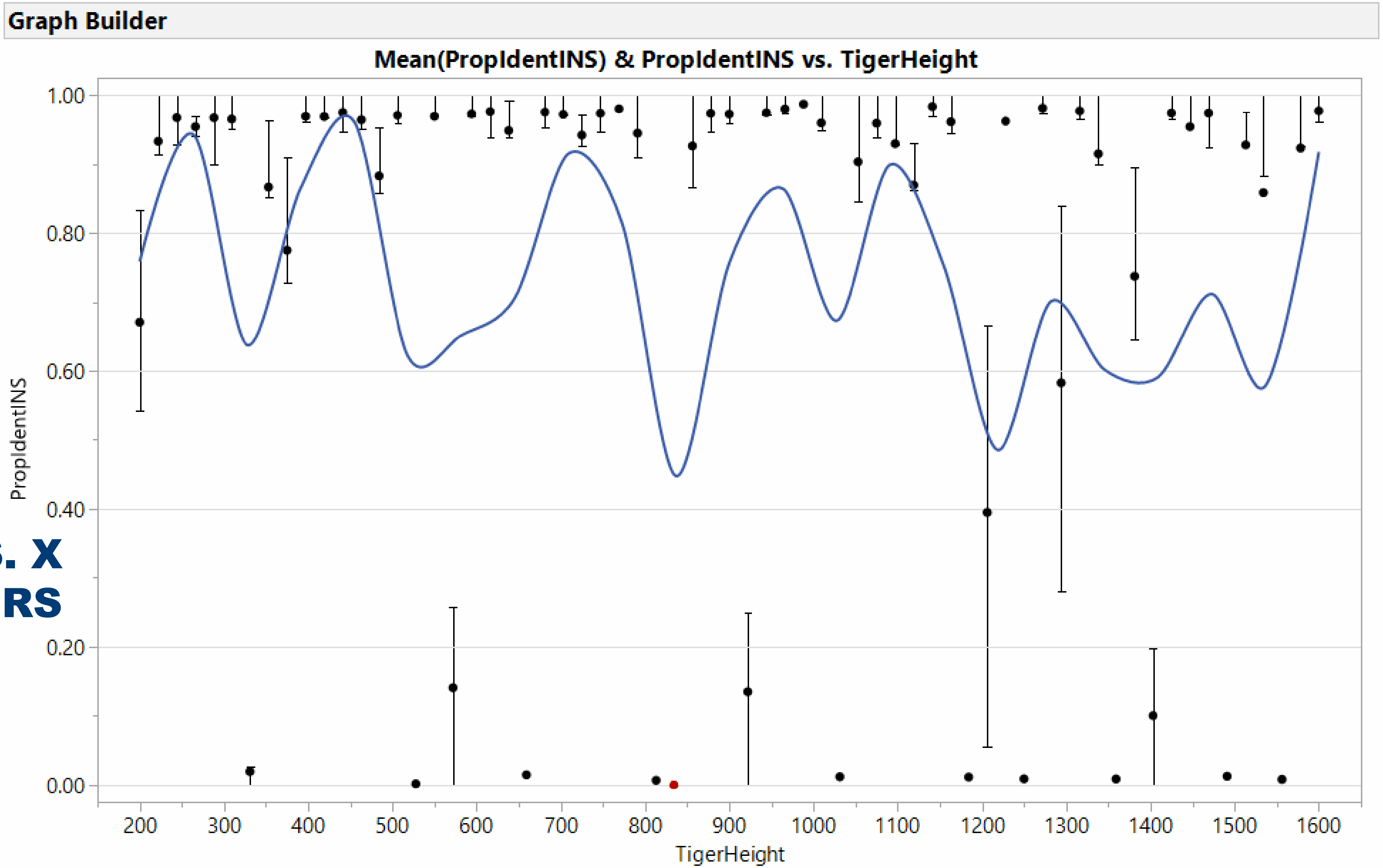
SPACE-FILLING DOE





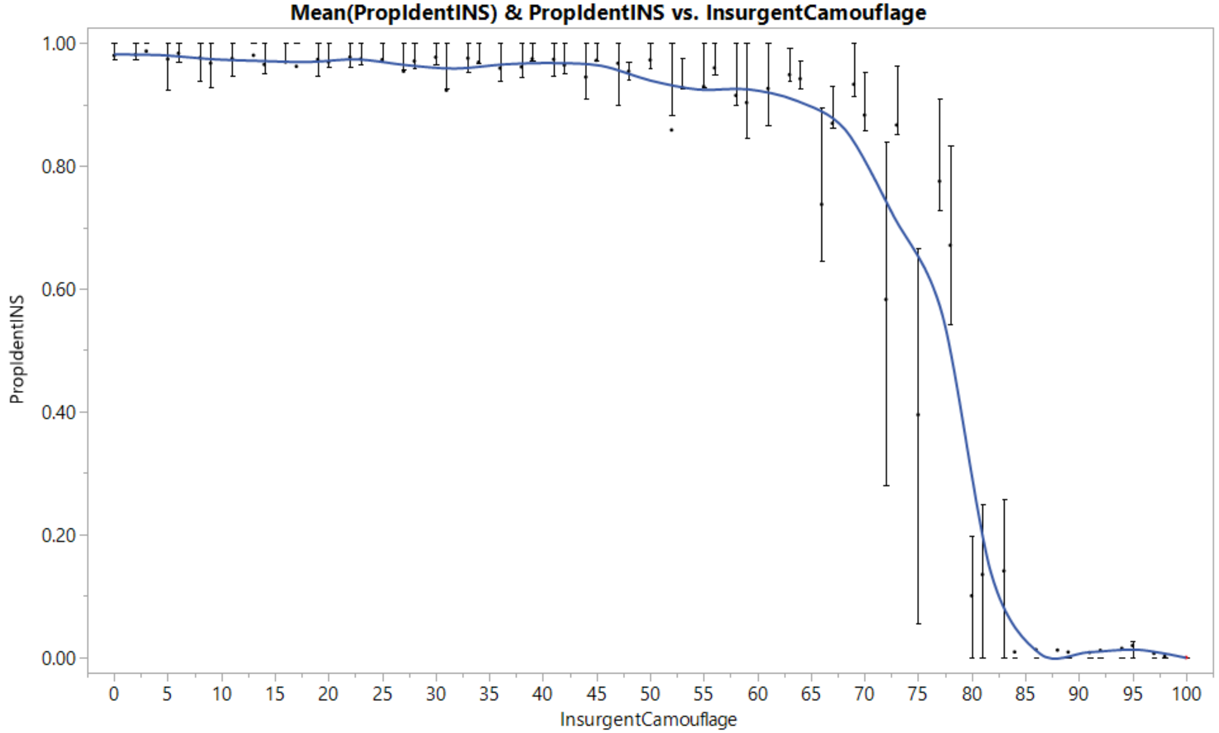


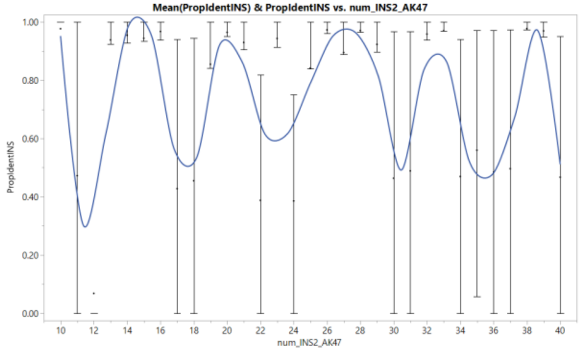
PROPIDENTINS VS. X FOR 6 FACTORS

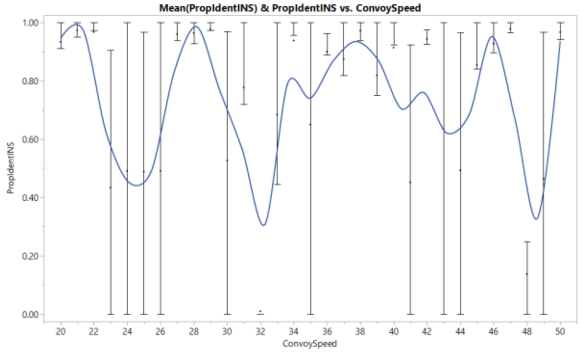


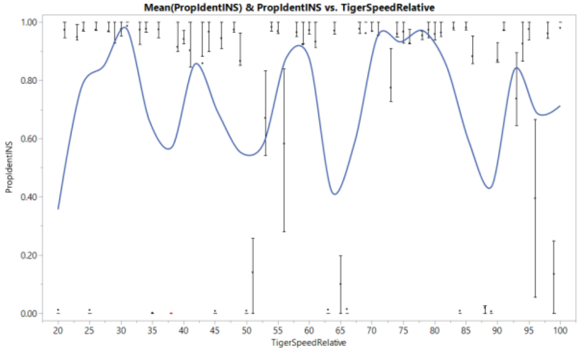
Each error bar is constructed using the upper and lower quartiles.

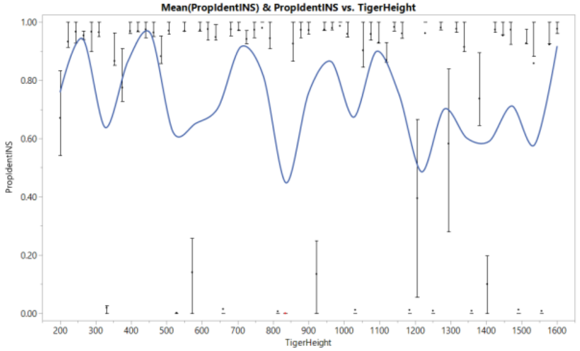
PROPIDENTINS VS. X FOR 6 FACTORS

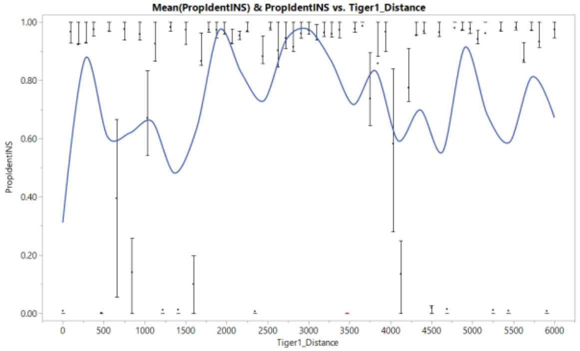




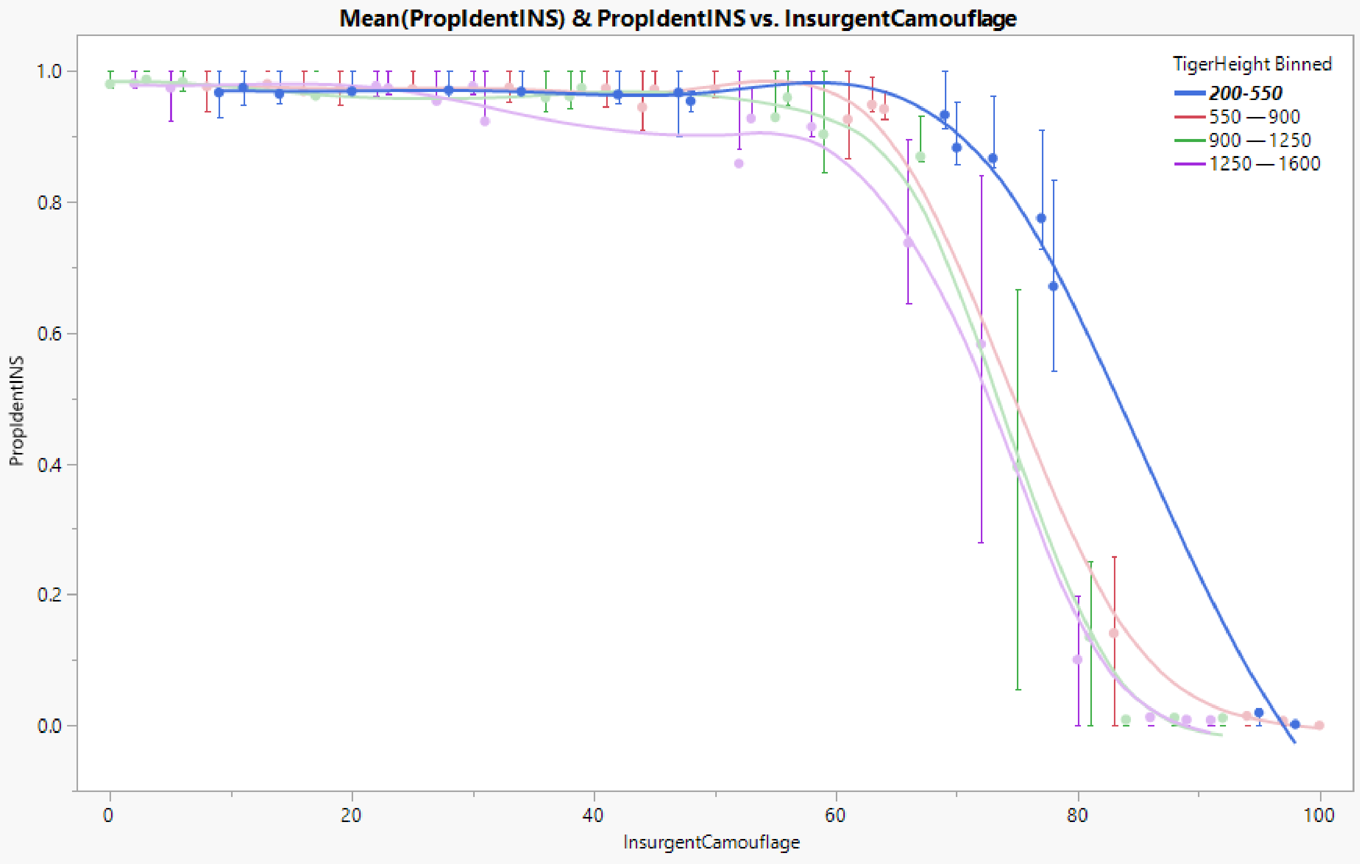


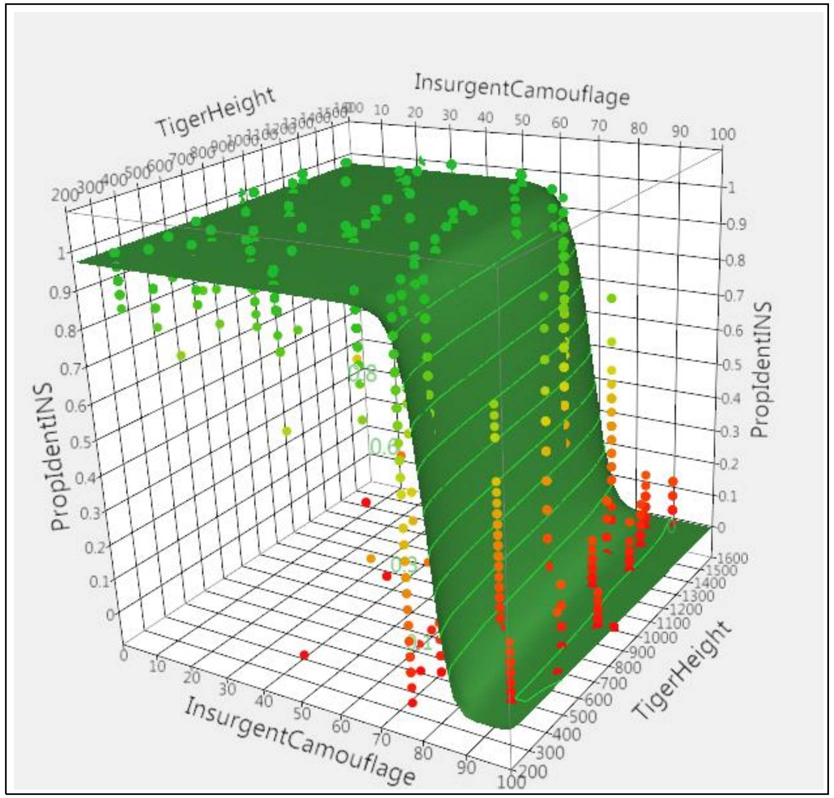






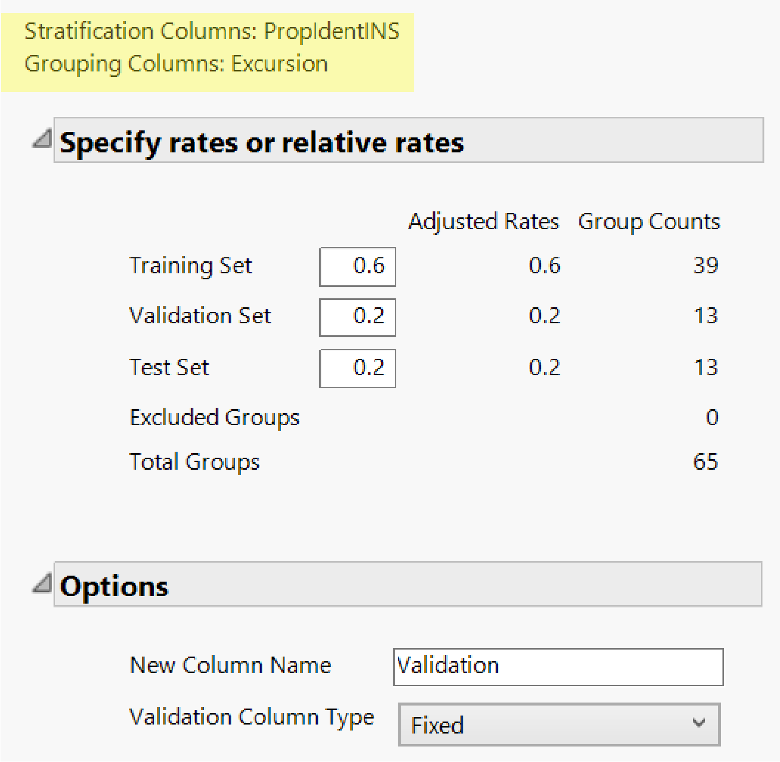
PROPIDENTINS VS. CAMOUFLAGE AT DIFFERENT HEIGHTS

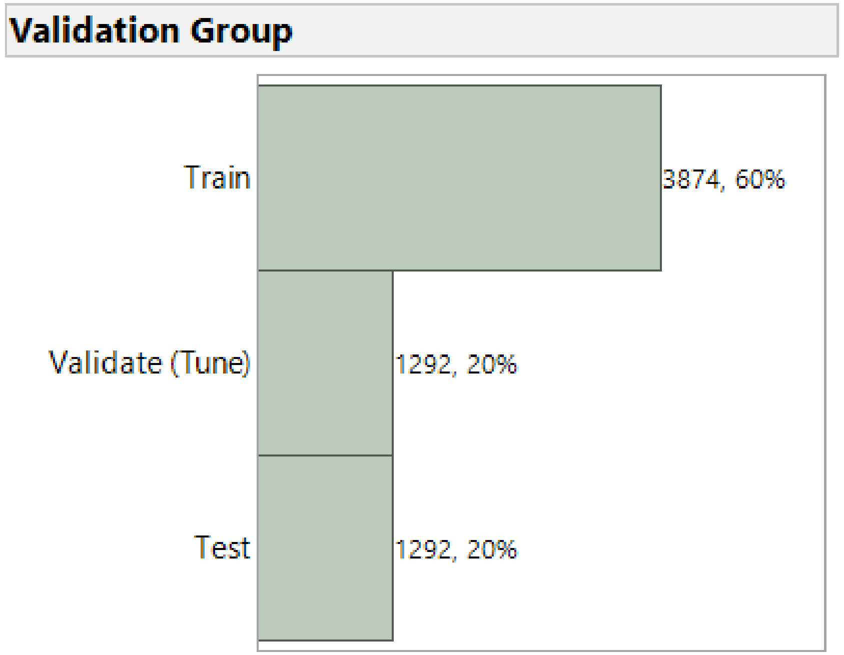




HONEST ASSESSMENT APPROACH
USING TRAIN, VALIDATE (TUNE), AND TEST SUBSETS

Used in model selection and estimating its prediction error on new data



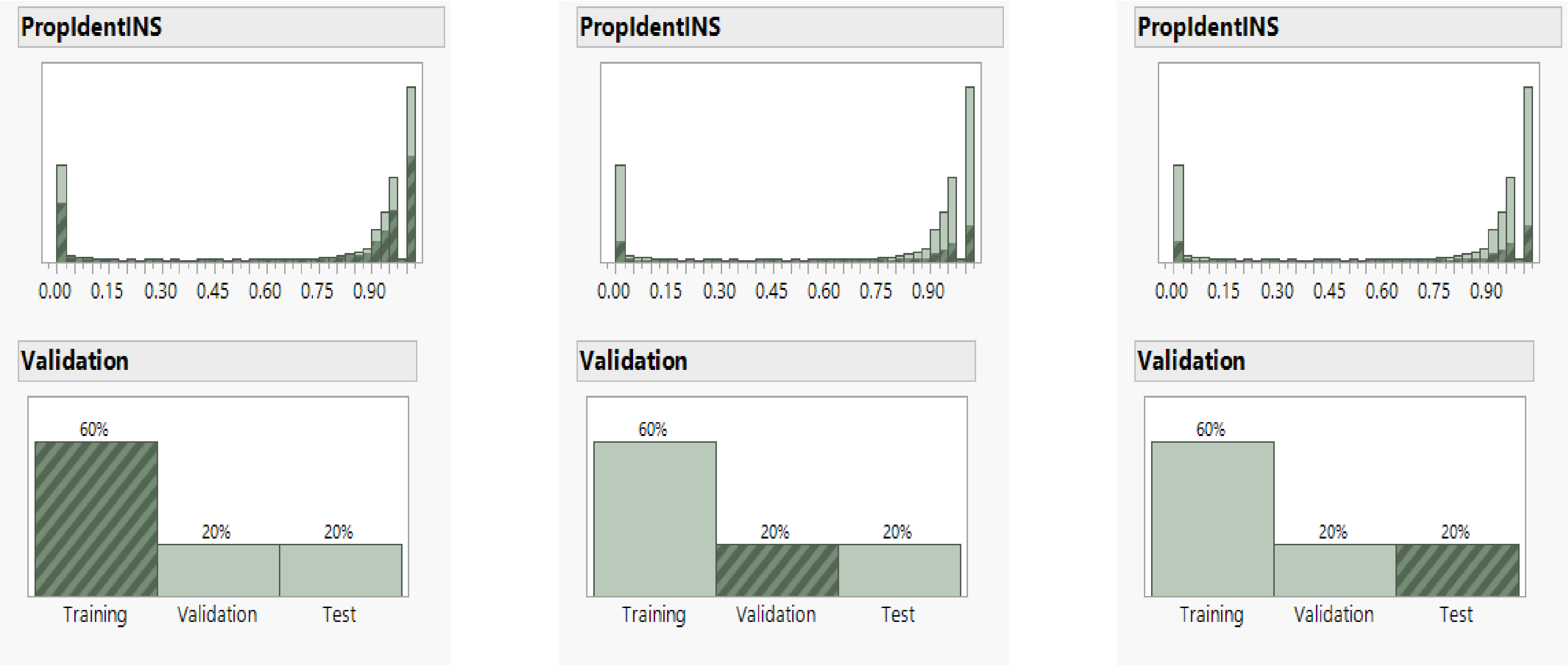


The Elements of Statistical Learning – Data Mining, Inference, and Prediction

Hastie, Tibshirani, and Friedman – 2001 (Chapter 7: Model Assessment and Selection)

HONEST ASSESSMENT APPROACH
USING TRAIN, VALIDATE (TUNE), AND TEST SUBSETS

NOTE: Same proportion of *PropIdentINS* in each Subset



The Elements of Statistical Learning – Data Mining, Inference, and Prediction

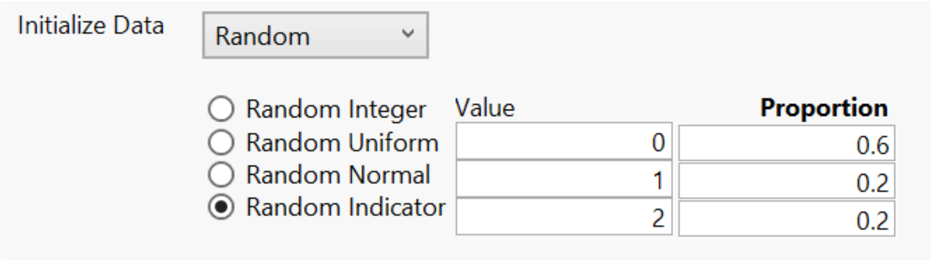
Hastie, Tibshirani, and Friedman – 2001

(Chapter 7: Model Assessment and Selection)

HONEST ASSESSMENT APPROACH USING TRAIN, VALIDATE (TUNE), AND TEST SUBSETS

Stratified Data Partitioning Add-in available for JMP (courtesy of the “”Data Doctor”)

Also, in base JMP:
Initialize Data Randomly in a new column (no stratification)



brady_brady STAFF

Stratified Data Partitioning (with balancing options) add-in.

Created: DEC 24, 2014 9:34 AM | Last Modified: NOV 27, 2017 1:05 PM

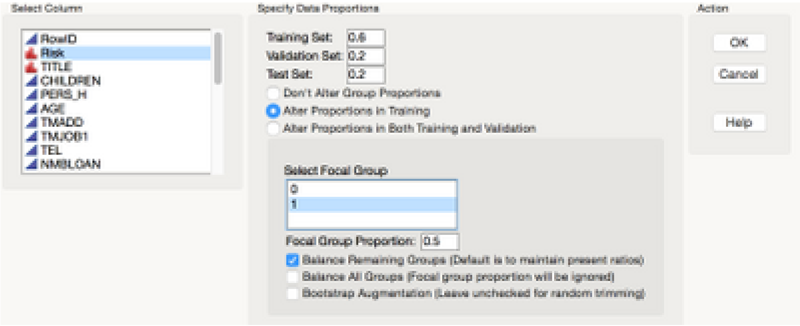
Stratified Split Balanced.jmpaddin

Stratified Data Partitioning Instructions rev3.pdf

This add-in allows the user to split a dataset into train/validate/test partitions. It includes options for rebalancing the proportions of the output data set's strata variable levels in relation to a focal group. This feature is useful, for example, in oversampling an event that is rare in the original data.

Instructions for using the add-in are attached.

Updated 3/23/2016: Includes additional balancing options.
Updated 9/1/2016: Bug fixes (related to an error when running the add-in)
Updated 9/2/2016: Added instructions (attached pdf)
Updated 11/27/2017: Uploaded revised instructions (attached pdf)



HONEST ASSESSMENT APPROACH USING TRAIN, VALIDATE (TUNE), AND TEST SUBSETS

michael_jmp STAFF

Imbalanced Classification Add-In

Created: **SEP 10, 2020 04:21 PM** | Last Modified: **MAR 26, 2021 08:48 AM**

Imbalanced Classification Version 2.jmpaddin

The Imbalanced Classification add-in features sampling techniques that attempt to impose a more balanced distribution between the two classes. The sampling techniques include the synthetic minority oversampling technique (SMOTE), Tomek links, and a combination of the two, as well as some basic sampling approaches. The Tomek Sampling, SMOTE Observations, and SMOTE plus Tomek options enable you to apply these sampling techniques on their own to support your specific modeling efforts.

The comprehensive Evaluate Models option, which requires JMP Pro, enables you to fit models using various sampling methods and compare them on a test set to select thresholds using Precision-Recall, ROC, and Cumulative Gains curves, as well as other measures of classification accuracy. The other three options do not fit models, but rather enable you to apply the Tomek, SMOTE, and SMOTE plus Tomek sampling schemes to your own data.

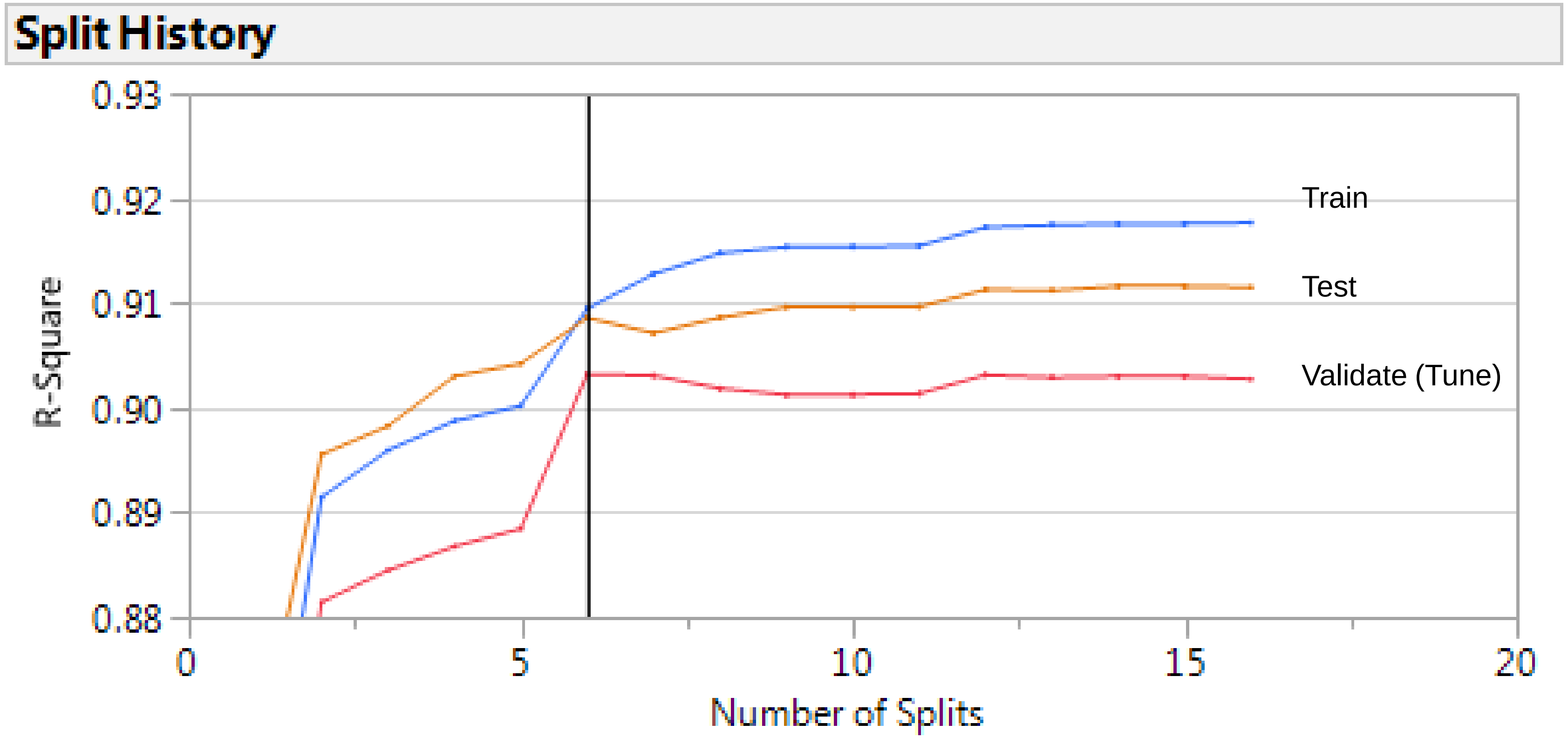
The SMOTE, Tomek, and combined SMOTE and Tomek sampling techniques use the concept of nearest neighbors. The add-in uses Gower distance as its distance metric, which allows for continuous, nominal, and ordinal predictors. These options do not require JMP Pro.

Note: All options require JMP version 15.2 or higher. Excluded rows and rows with missing response values are ignored by the add-in.

Version 2, released 3/25/2021, supports JMP 16 and improves the handling of rows with missing values for all predictors.

Discovery Summit Video Link

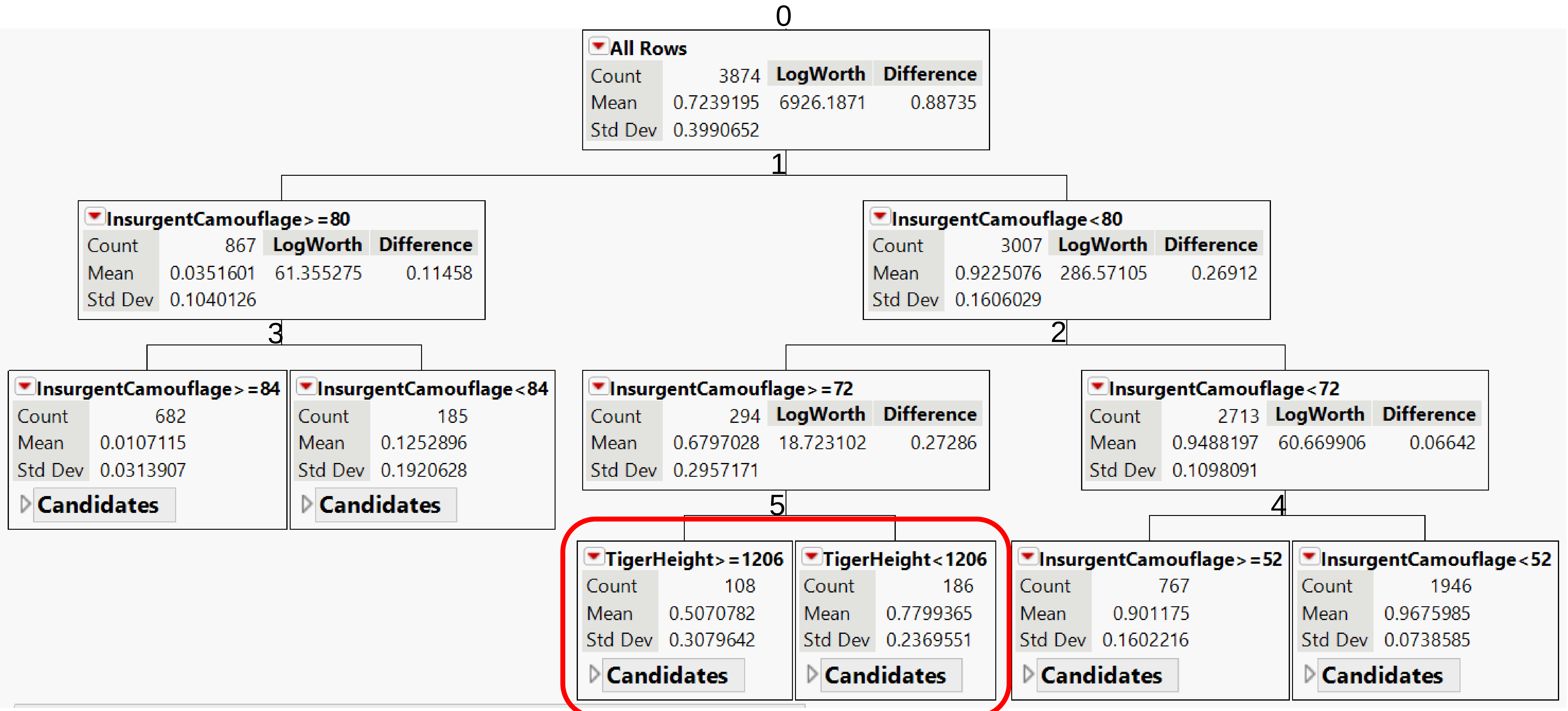
R-SQUARE VS. NUMBER OF SPLITS (FOR 1 RANDOM TVT)



Validation Data in Red
Test Data in Orange

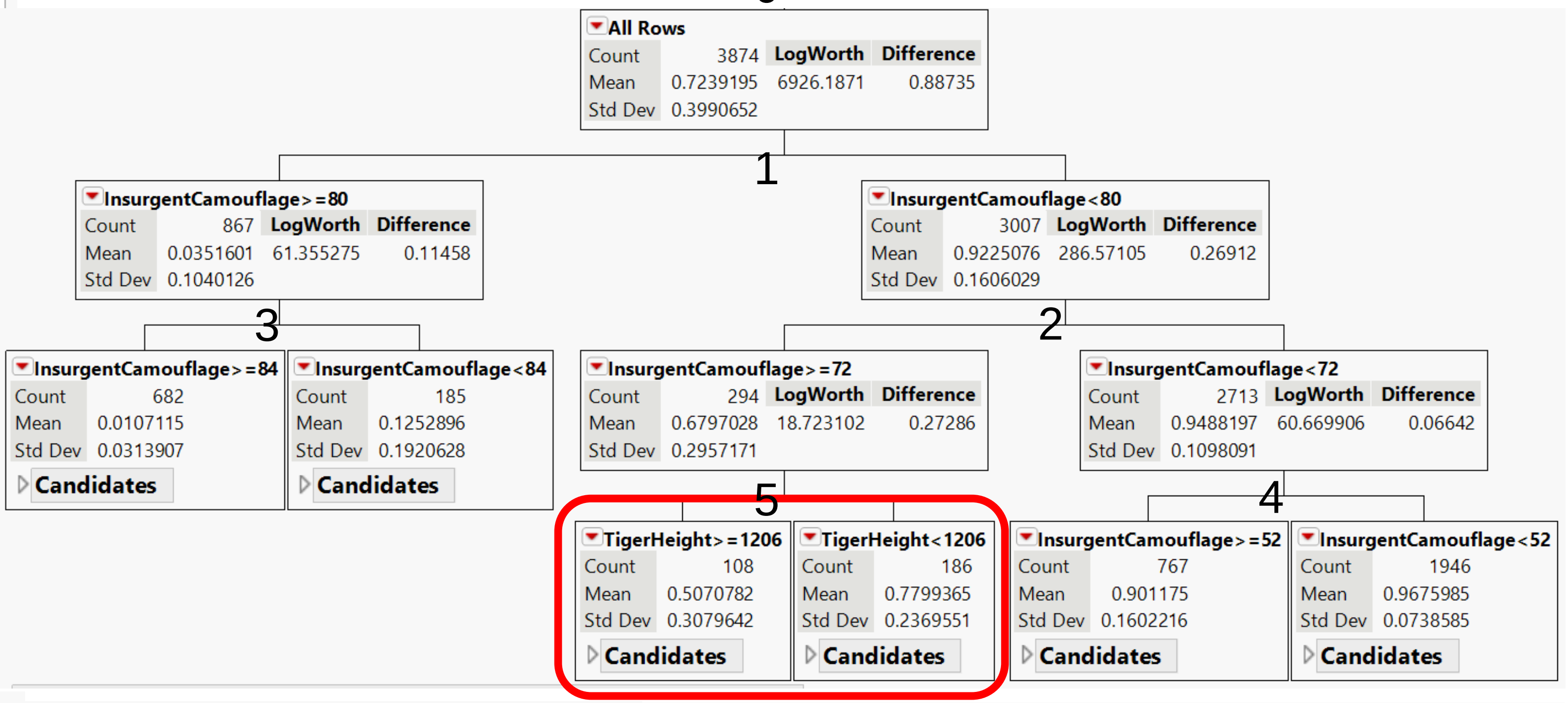
DECISION TREE

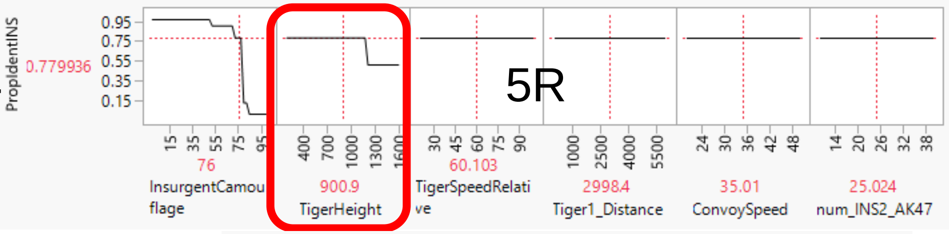
Each split finds the cut point among all factors that creates the biggest difference in the means of the two partitions of the data

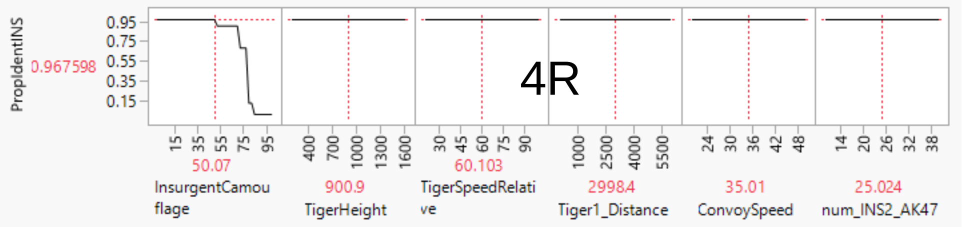


DECISION TREE

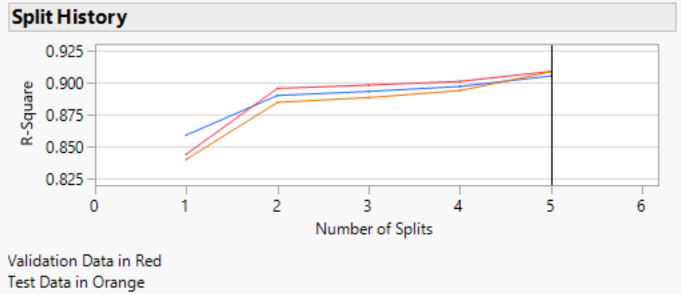
Each split finds the cut point among all factors that creates the biggest difference in the means of the two partitions of the data







Can be interpreted as a series of nested "If" statements

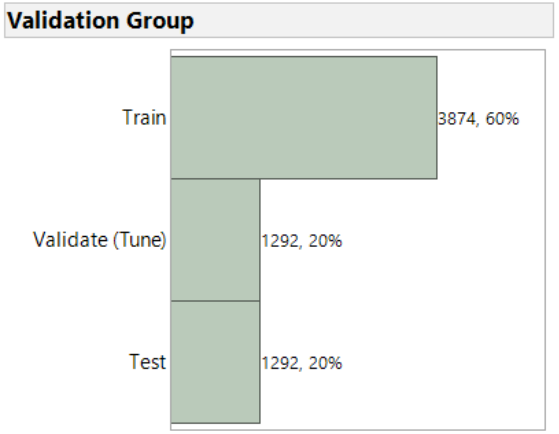


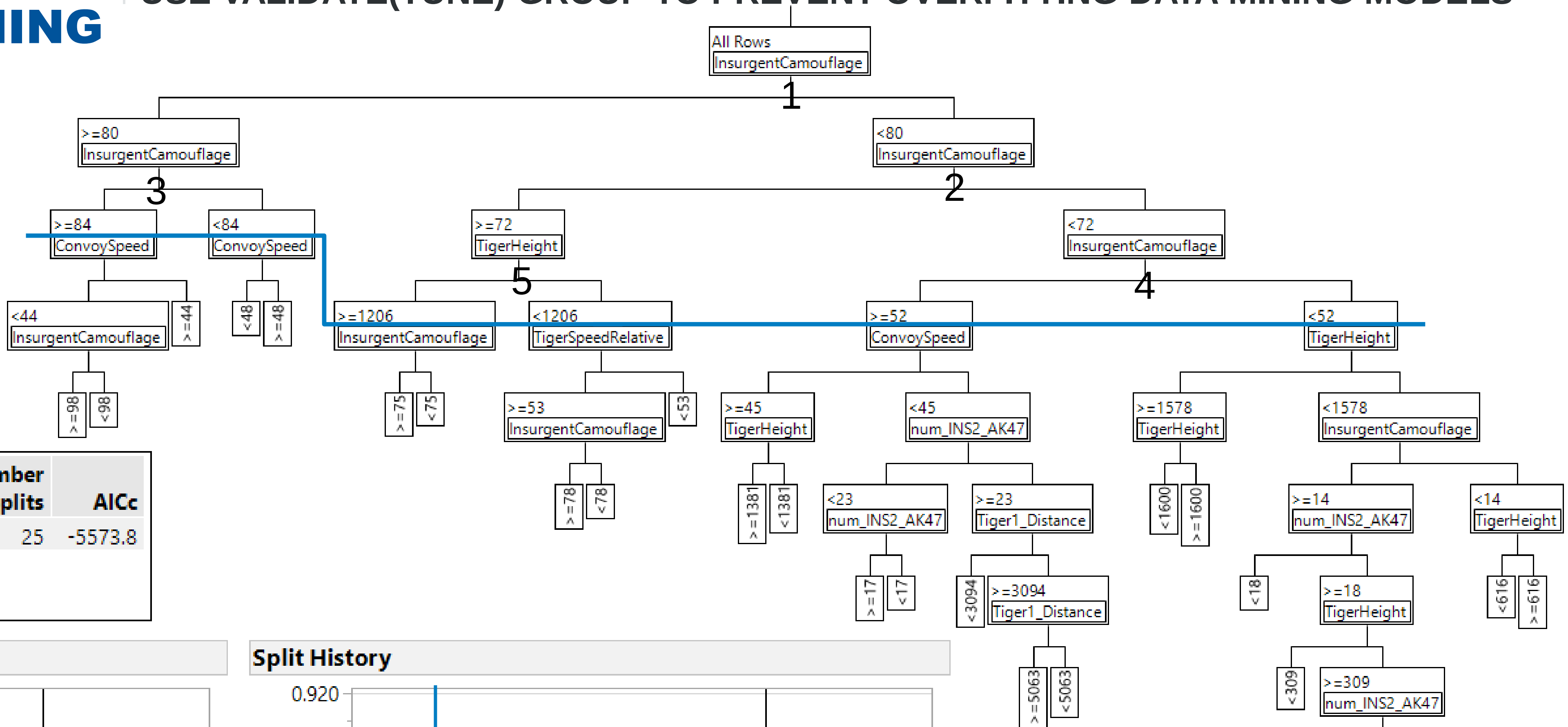
Column Contributions

Term	Number of Splits	SS	Portion
InsurgentCamouflage	4	553.432098	0.9909
TigerHeight	1	5.08702203	0.0091
TigerSpeedRelative	0	0	0.0000
Tiger1_Distance	0	0	0.0000
ConvoySpeed	0	0	0.0000
num_INS2_AK47	0	0	0.0000

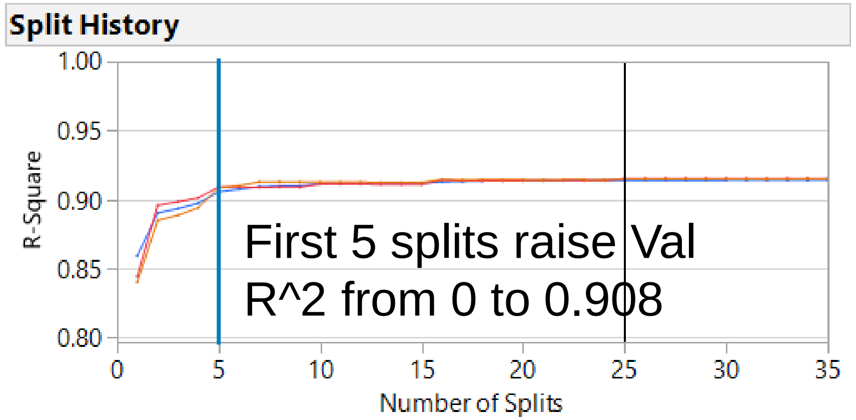
HONEST ASSESSMENT WHEN DATA MINING

SUBSET DATA TO CREATE *TRAIN*, *VALIDATE*(*TUNE*), & *TEST* GROUPS
USE VALIDATE(TUNE) GROUP TO PREVENT OVERFITTING DATA MINING MODELS

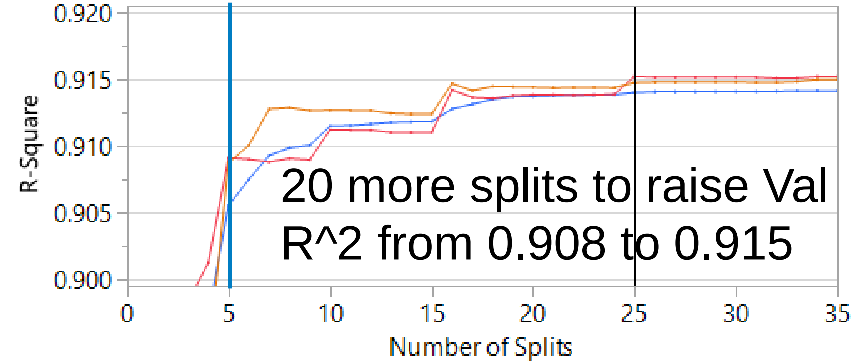




	RSquare	RMSE	N	Number of Splits	AICc
Training	0.914	0.1170276	3874	25	-5573.8
Validation	0.915	0.1132339	1292		
Test	0.915	0.1147605	1292		



Validation Data in Red
Test Data in Orange



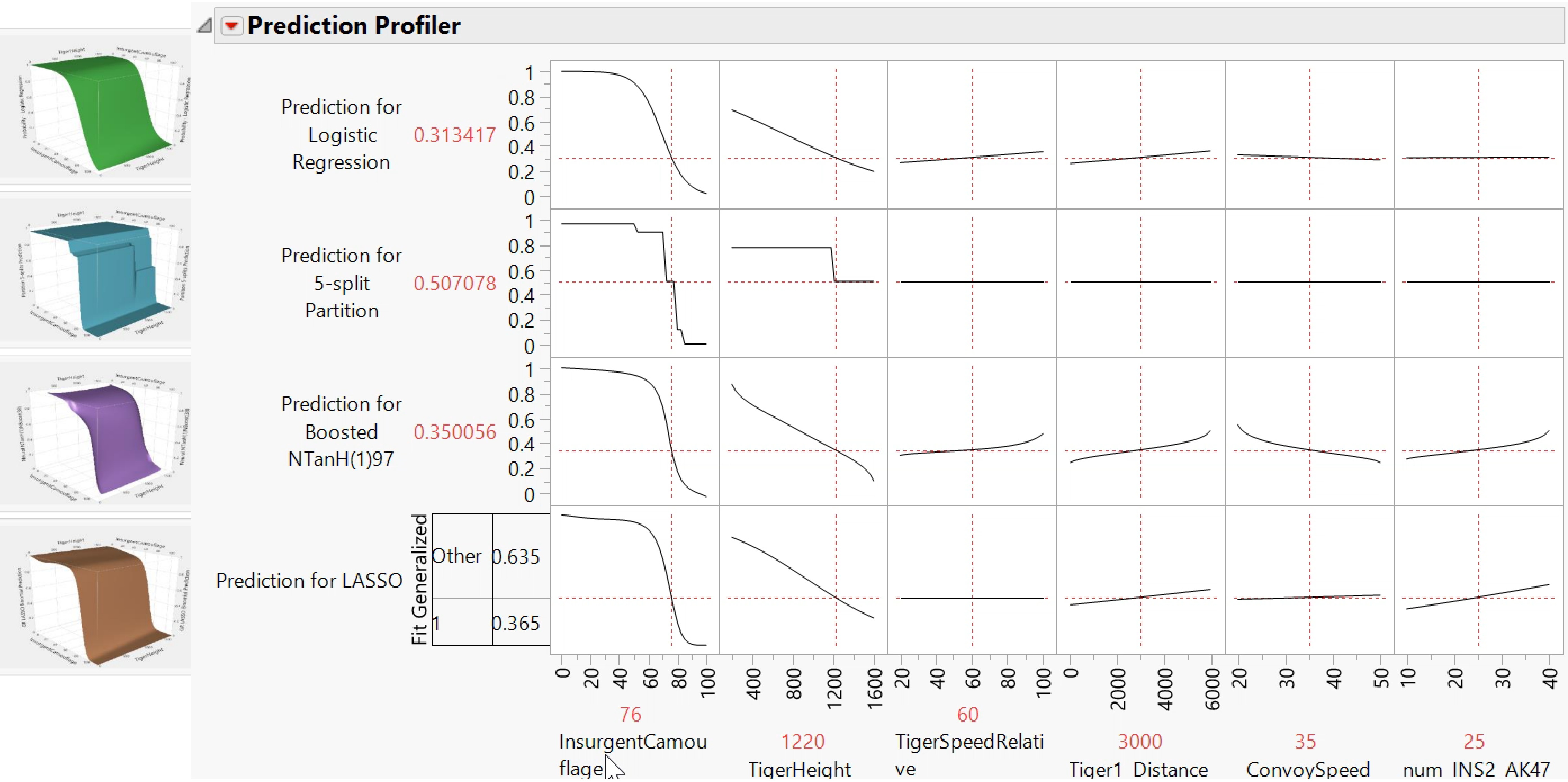
Validation Data in Red
Test Data in Orange

Column Contributions

Term	Number of Splits	SS	Portion
InsurgentCamouflage	9	555.084982	0.9847
TigerHeight	6	6.46096421	0.0115
ConvoySpeed	4	1.45893941	0.0026
num_INS2_AK47	4	0.66588349	0.0012
Tiger1_Distance	2	0.06006294	0.0001
TigerSpeedRelative	0	0	0.0000

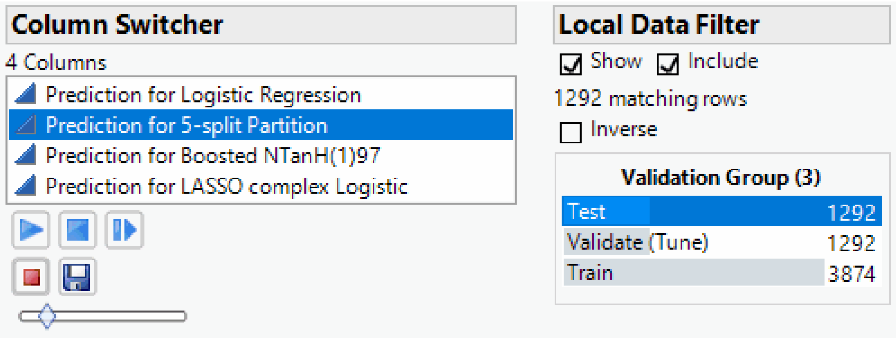
COMPARE SEVERAL MODELS

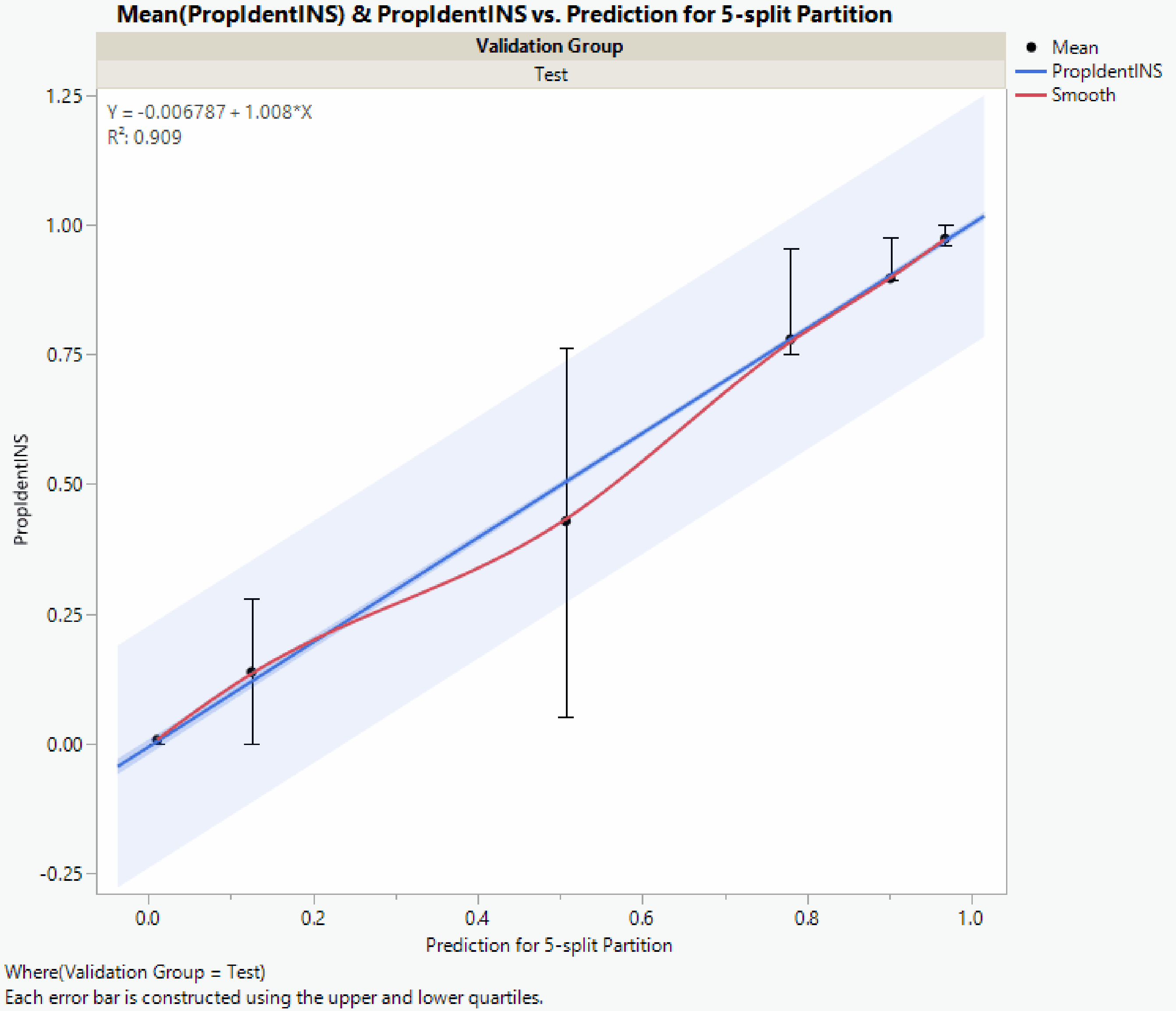
Logistic Regression, Partition with 5-Splits, Neural Network, & LASSO Binomial



R^2
0.876
0.908
0.912
0.916

ACTUAL VS. PREDICTED PLOTS FOR ***TEST DATA ONLY***

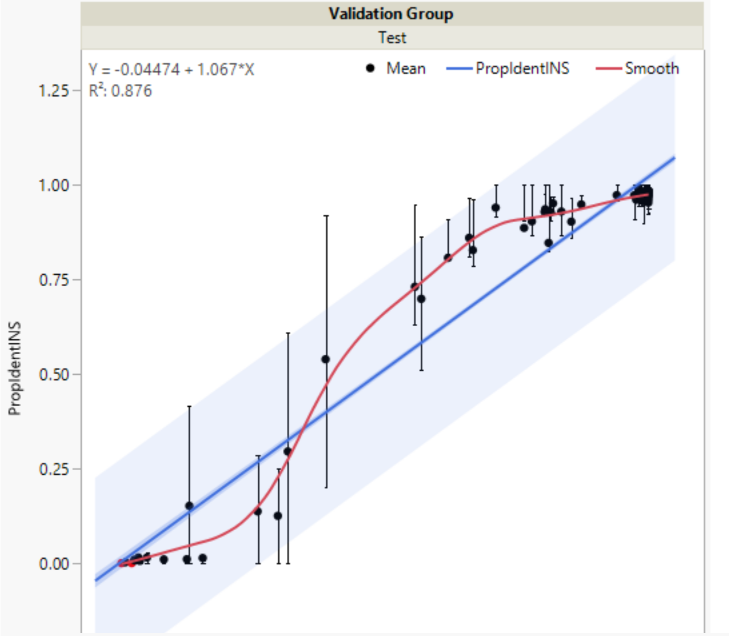




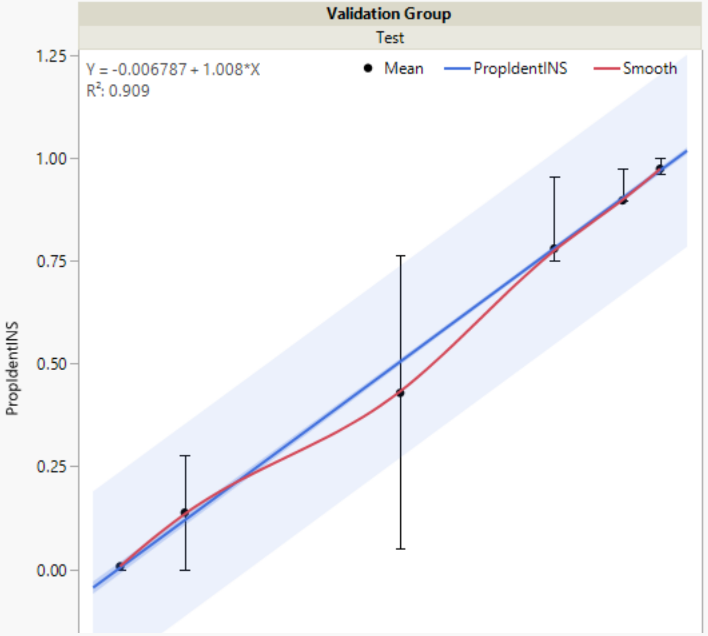
Four Models

1. Logistic Regression
2. Partition with 5-Splits
3. Neural Network
4. LASSO (Binomial Distribution)

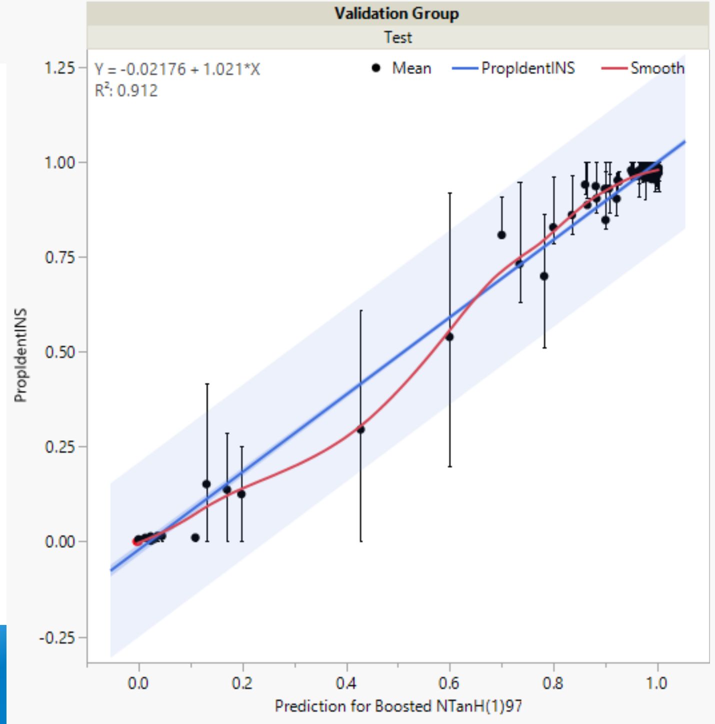
Mean(PropIdentINS) & PropIdentINS vs. Prediction for Logistic Regression



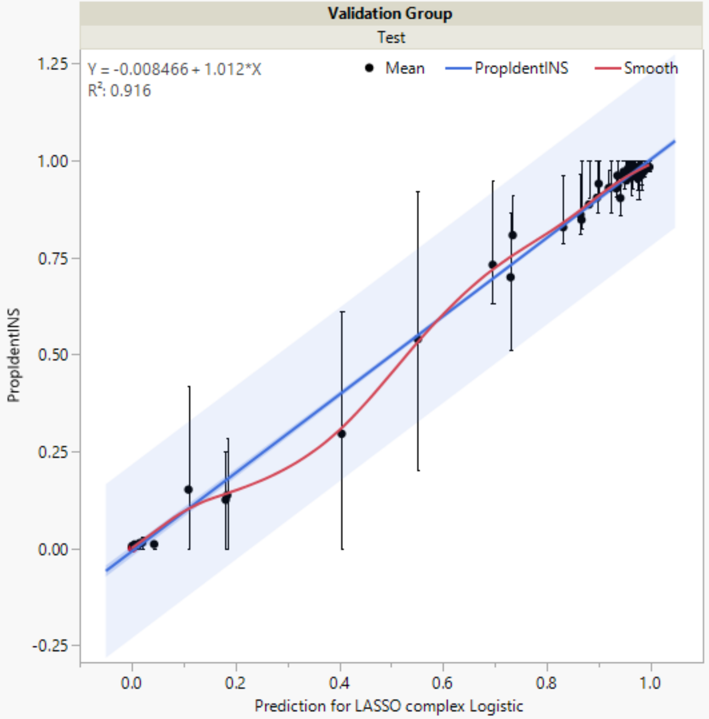
Mean(PropIdentINS) & PropIdentINS vs. Prediction for 5-split Partition



Mean(PropIdentINS) & PropIdentINS vs. Prediction for Boosted NTanH(1)97



Mean(PropIdentINS) & PropIdentINS vs. Prediction for LASSO complex Logistic

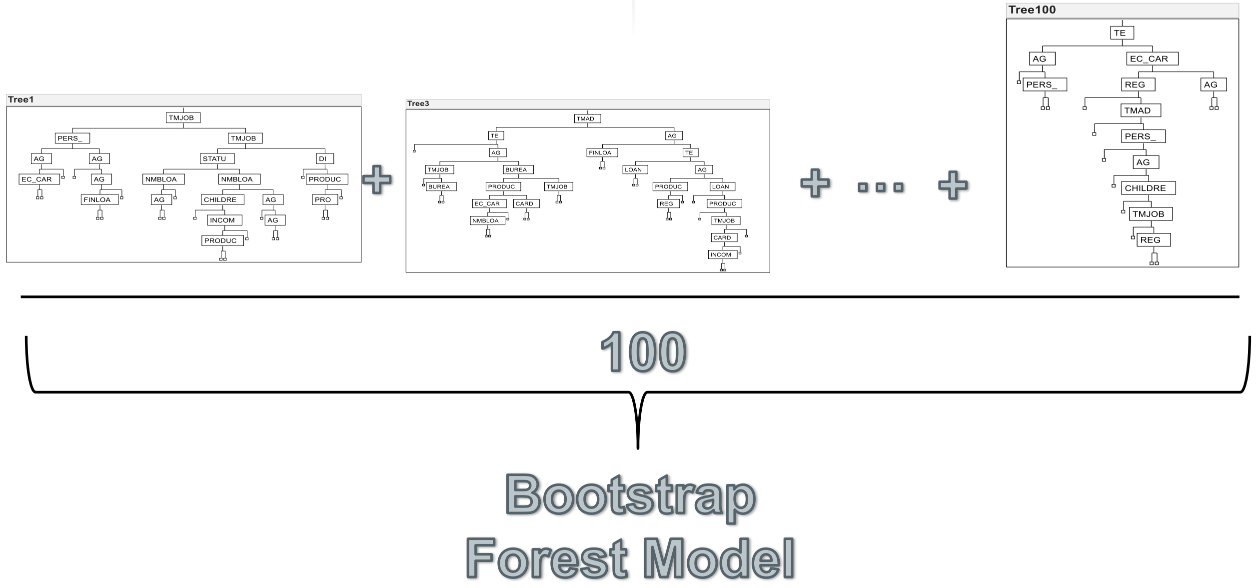


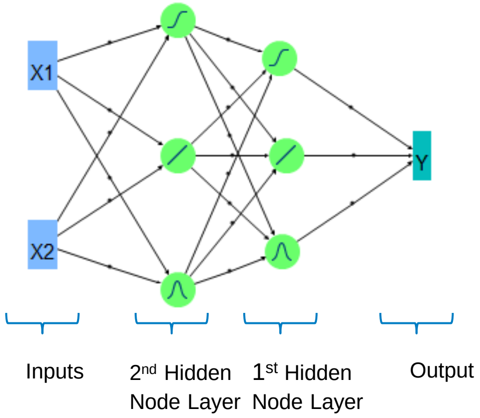
ACTUAL VS. PREDICTED PLOTS FOR ***TEST DATA ONLY***

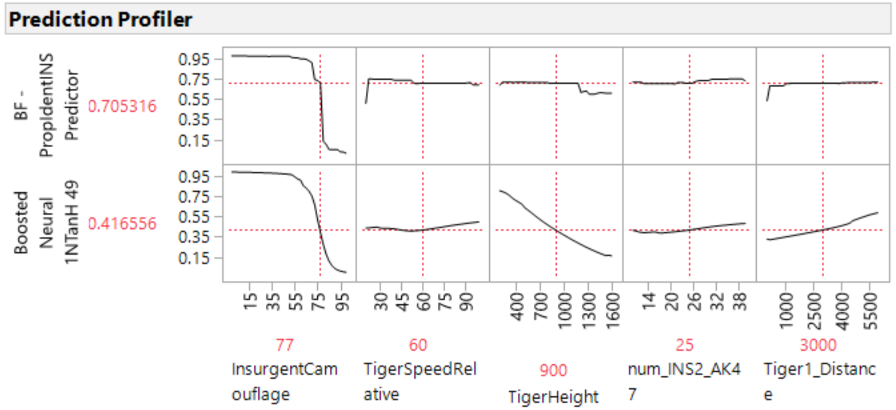
LOGISTIC REGRESSION
PARTITION WITH 5-SPLITS
NEURAL NETWORK
LASSO (BINOMIAL DIST.)

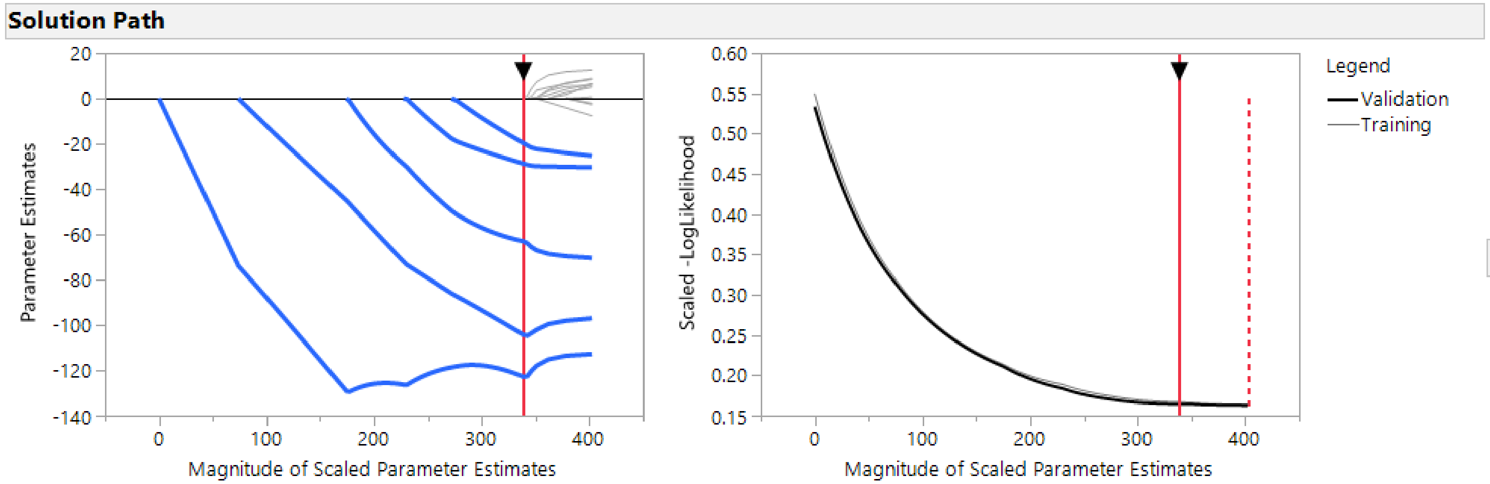
MACHINE LEARNING ROBUST STRATEGY

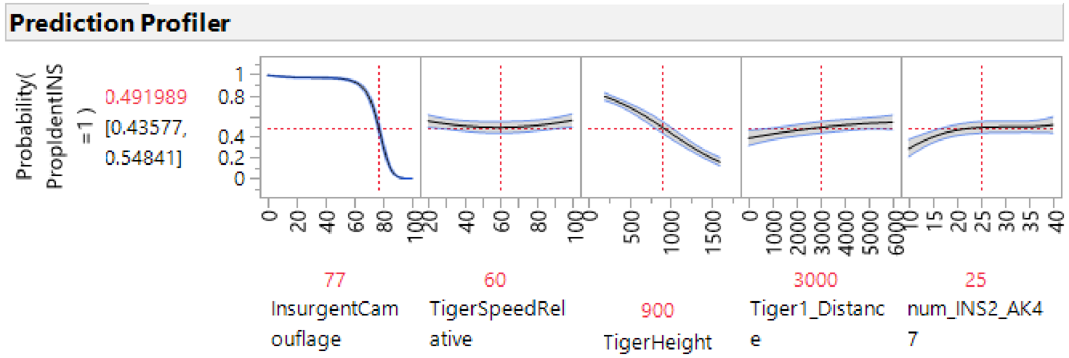
1) BOOTSTRAP FOREST DECISION TREE – DON'T MISS AN IMPORTANT VARIABLE
2) NEURAL NETWORK – OFTEN MOST FLEXIBLE & BEST PREDICTING MODEL
3) PENALIZED REGRESSION – MORE INTERPRETABLE MODEL + CONF. INTERVALS AND CAN BE NEARLY AS ACCURATE AS NEURAL NETWORK





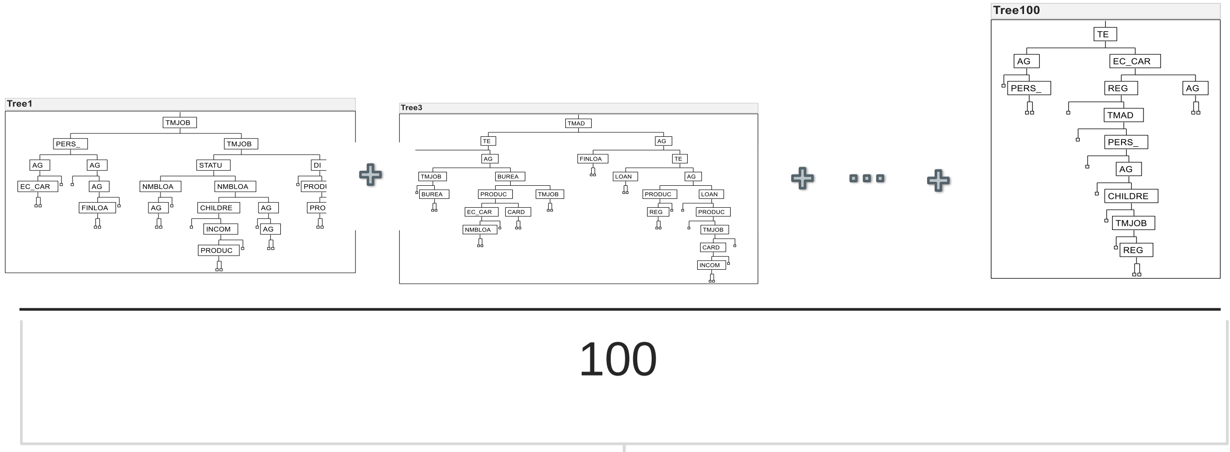




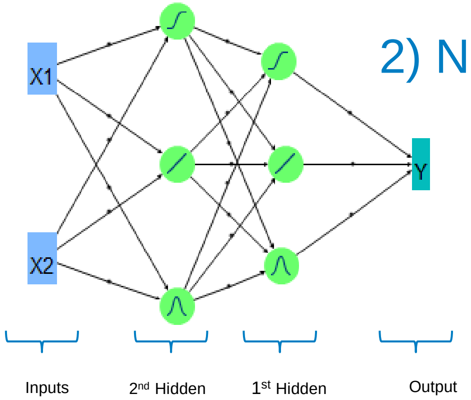


ROBUST STRATEGY FOR MACHINE LEARNING

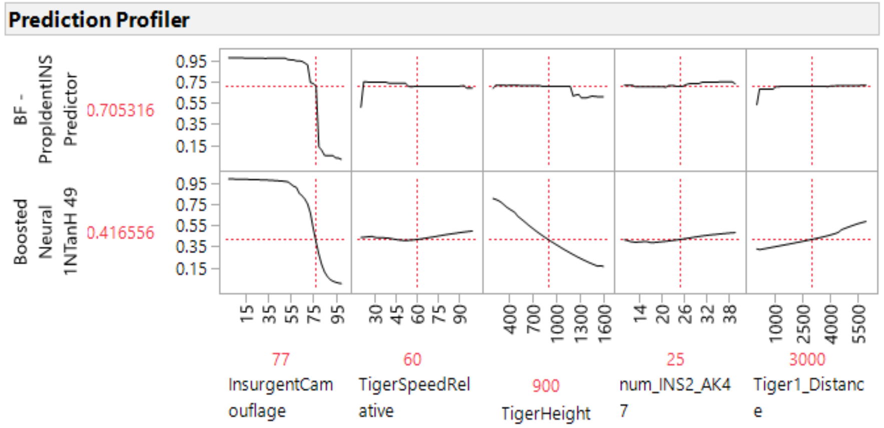
1) BOOTSTRAP FOREST DECISION TREE – DON'T MISS AN IMPORTANT VARIABLE
2) NEURAL NETWORK – OFTEN MOST FLEXIBLE & BEST PREDICTING MODEL
3) PENALIZED REGRESSION – MORE INTERPRETABLE MODEL + CONF. INTERVALS

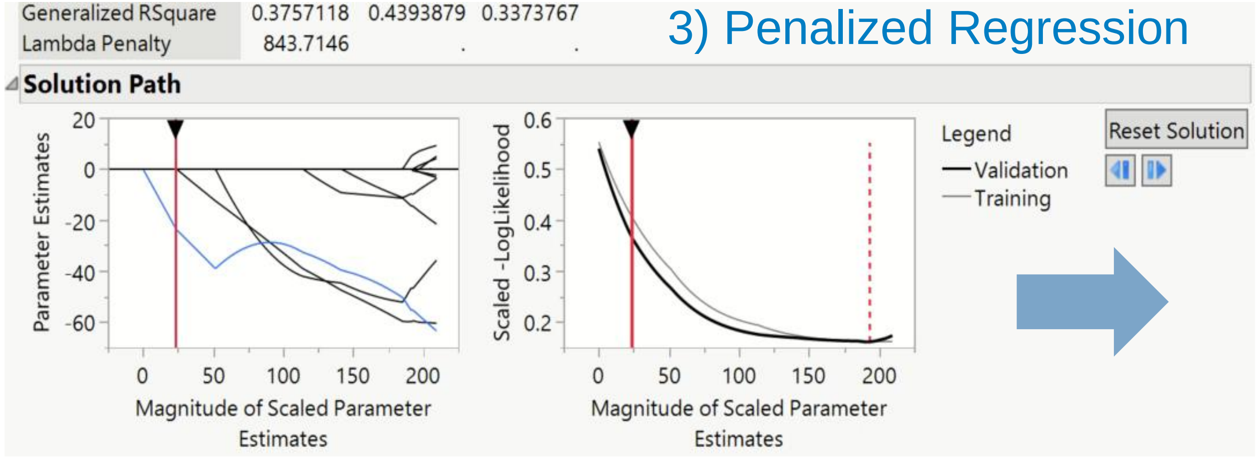


1) Bootstrap Forest

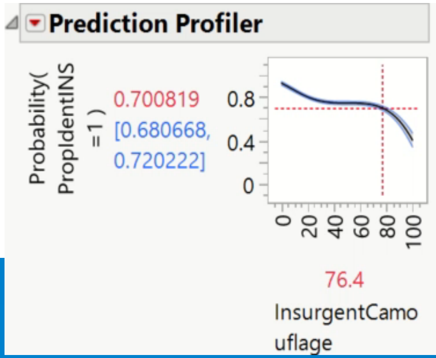


2) Neural Net





3) Penalized Regression



jmp *NOT just better prediction, but better understanding!*

BOOTSTRAP FOREST – VARIABLE SELECTION W/44 FACTORS

3 dummy factors created from random data

Column Contributions

Term	Number of Splits	G^2	Portion
service	450	10603400.8	0.2831
dst_bytes	382	5308498.33	0.1417
src_bytes	820	4771327.16	0.1274
count	337	2700247.28	0.0721
dst_host_srv_count	528	1990388.66	0.0531
dst_host_diff_srv_rate	415	1575488.06	0.0421
flag	168	1153015.42	0.0308
srv_count	238	1115688.05	0.0298
dst_host_serror_rate	175	1060259.19	0.0283
duration	276	991351.909	0.0265
dst_host_count	499	714300.159	0.0191
dst_host_same_src_port_rat	389	616742.634	0.0165
hot	159	535399.996	0.0143
same_srv_rate	103	422795.794	0.0113
dst_host_same_srv_rate	334	421699.768	0.0113
diff_srv_rate	145	382986.204	0.0102
serror_rate	65	365667.013	0.0098
dst_host_rerror_rate	233	318445.492	0.0085
dst_host_srv_serror_rate	117	308717.284	0.0082
logged_in	40	305603.637	0.0082
srv_serror_rate	30	219339.913	0.0059
root_shell	32	203921.266	0.0054
dst_host_srv_diff_host_rate	253	196905.011	0.0053
Random Uniform	228	195145.878	0.0052
dst_host_srv_rerror_rate	81	153228.513	0.0041
protocol_type	53	152857.046	0.0041
is_guest_login	12	137886.036	0.0037
Random Normal	194	110253.474	0.0029
num_compromised	39	76703.4706	0.0020
num_file_creations	20	75279.6937	0.0020
wrong_fragment	29	72313.7688	0.0019
rerror_rate	45	59525.1111	0.0016
num_root	23	41990.5367	0.0011
Random Integer	146	21117.3276	0.0006
srv_diff_host_rate	33	17448.0232	0.0005
num_failed_logins	7	17407.5895	0.0005
srv_rerror_rate	30	16080.2873	0.0004
num_access_files	11	11528.8834	0.0003
num_shells	11	8067.77994	0.0002
urgent	4	3131.15585	0.0001
su_attempted	1	42.7170189	0.0000
land	0	0	0.0000
num_outbound_cmds	0	0	0.0000
is_host_login	0	0	0.0000

Column Contributions

Term	Number of Splits	G^2	Portion
service	450	10603400.8	0.2831
dst_bytes	382	5308498.33	0.1417
src_bytes	820	4771327.16	0.1274
count	337	2700247.28	0.0721
dst_host_srv_count	528	1990388.66	0.0531
dst_host_diff_srv_rate	415	1575488.06	0.0421
flag	168	1153015.42	0.0308
srv_count	238	1115688.05	0.0298
dst_host_serror_rate	175	1060259.19	0.0283
duration	276	991351.909	0.0265
dst_host_count	499	714300.159	0.0191
dst_host_same_src_port_rat	389	616742.634	0.0165
hot	159	535399.996	0.0143
same_srv_rate	103	422795.794	0.0113
dst_host_same_srv_rate	334	421699.768	0.0113
diff_srv_rate	145	382986.204	0.0102

Top 11 of 44

Model Validation-Set Summaries

The fit below was the best of these models fit.

N Terms	N Trees	Entropy RSquare	Misclassification Rate	Avg -Log p	RMS Error	Avg Abs Error
11	200	0.9786	0.0040	0.0336	0.0856	0.0279
14	53	0.9811	0.0040	0.0297	0.0816	0.0243
18	48	0.9831	0.0039	0.0265	0.0770	0.0215

FAST VARIABLE SELECTION WITH 200 CONT. & 50 CAT. FACTORS & 12,000 ROWS
BOOTSTRAP FOREST (LEFT) & PREDICTOR SCREENING (RIGHT)

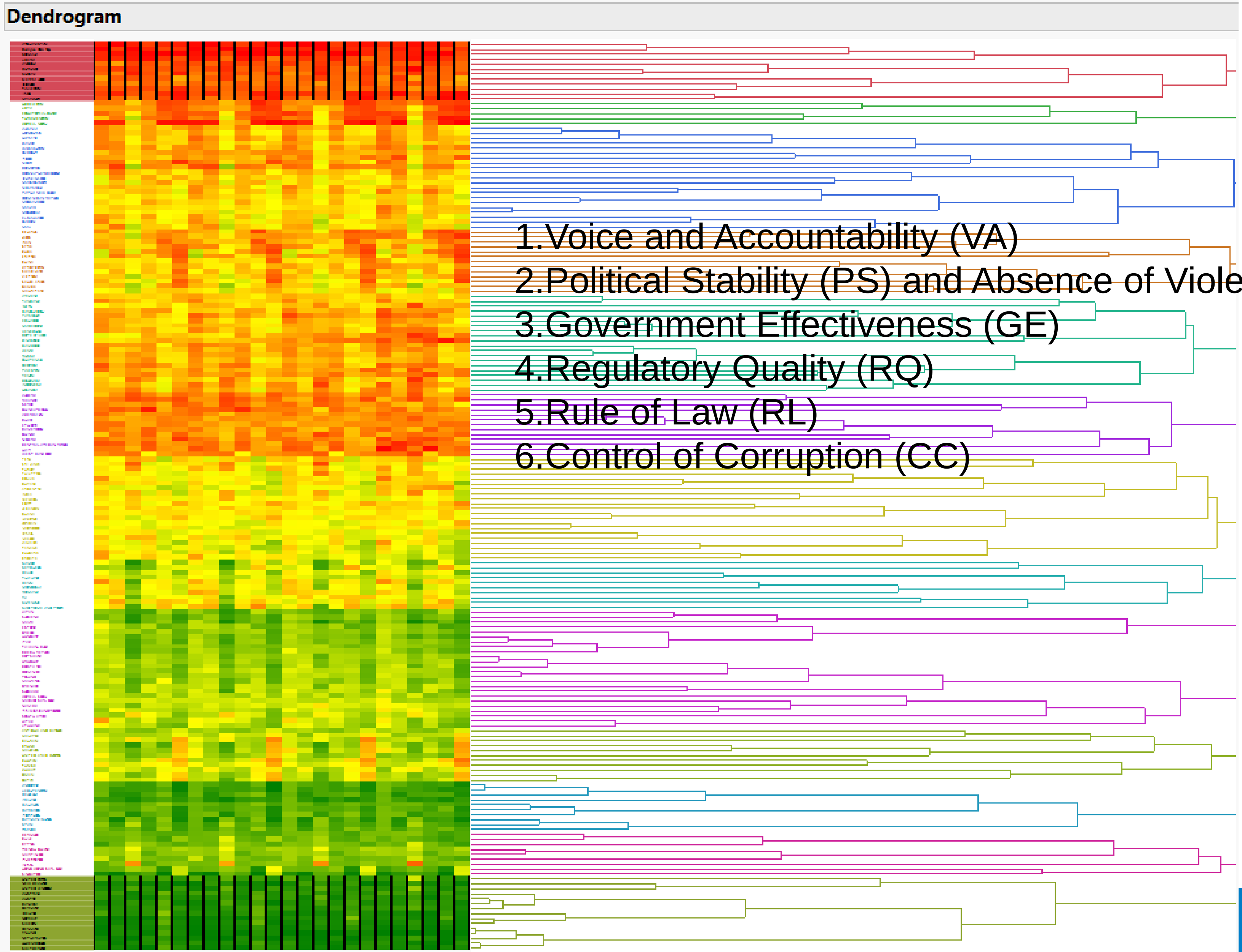
Column Contributions

Term	Number of Splits	SS		Portion
x.4	1616	32177.8441		0.3055
x.2	1203	17821.6507		0.1692
x.1	1151	17656.8273		0.1676
x.5	918	7450.24401		0.0707
x.3	940	4837.15111		0.0459
cat.208	266	317.387862		0.0030
cat.203	282	316.048361		0.0030
cat.201	279	313.582113		0.0030
cat.232	279	303.452344		0.0029
cat.233	264	300.630441		0.0029
cat.228	257	298.163627		0.0028
cat.206	254	297.002193		0.0028
cat.204	257	296.604953		0.0028
cat.246	268	294.989348		0.0028
cat.207	260	294.710682		0.0028
cat.226	247	291.120065		0.0028
cat.216	252	286.631695		0.0027
cat.241	248	283.205332		0.0027
cat.249	257	282.167316		0.0027

Predictor Screening

	y			
Predictor	Contribution	Portion		Rank
x.4	42854.3	0.3864		1
x.1	24757.2	0.2233		2
x.2	24023.8	0.2166		3
x.5	9085.2	0.0819		4
x.3	5268.8	0.0475		5
cat.227	66.3	0.0006		6
cat.228	63.1	0.0006		7
cat.236	62.7	0.0006		8
cat.212	61.0	0.0006		9
cat.215	59.7	0.0005		10
cat.246	58.8	0.0005		11
cat.223	58.7	0.0005		12
cat.206	57.0	0.0005		13
cat.201	56.7	0.0005		14
cat.241	56.4	0.0005		15
cat.232	54.8	0.0005		16
cat.239	53.0	0.0005		17
cat.231	52.9	0.0005		18
cat.216	52.2	0.0005		19
cat.240	52.0	0.0005		20

UNSUPERVISED ML CLUSTERING OF DATA



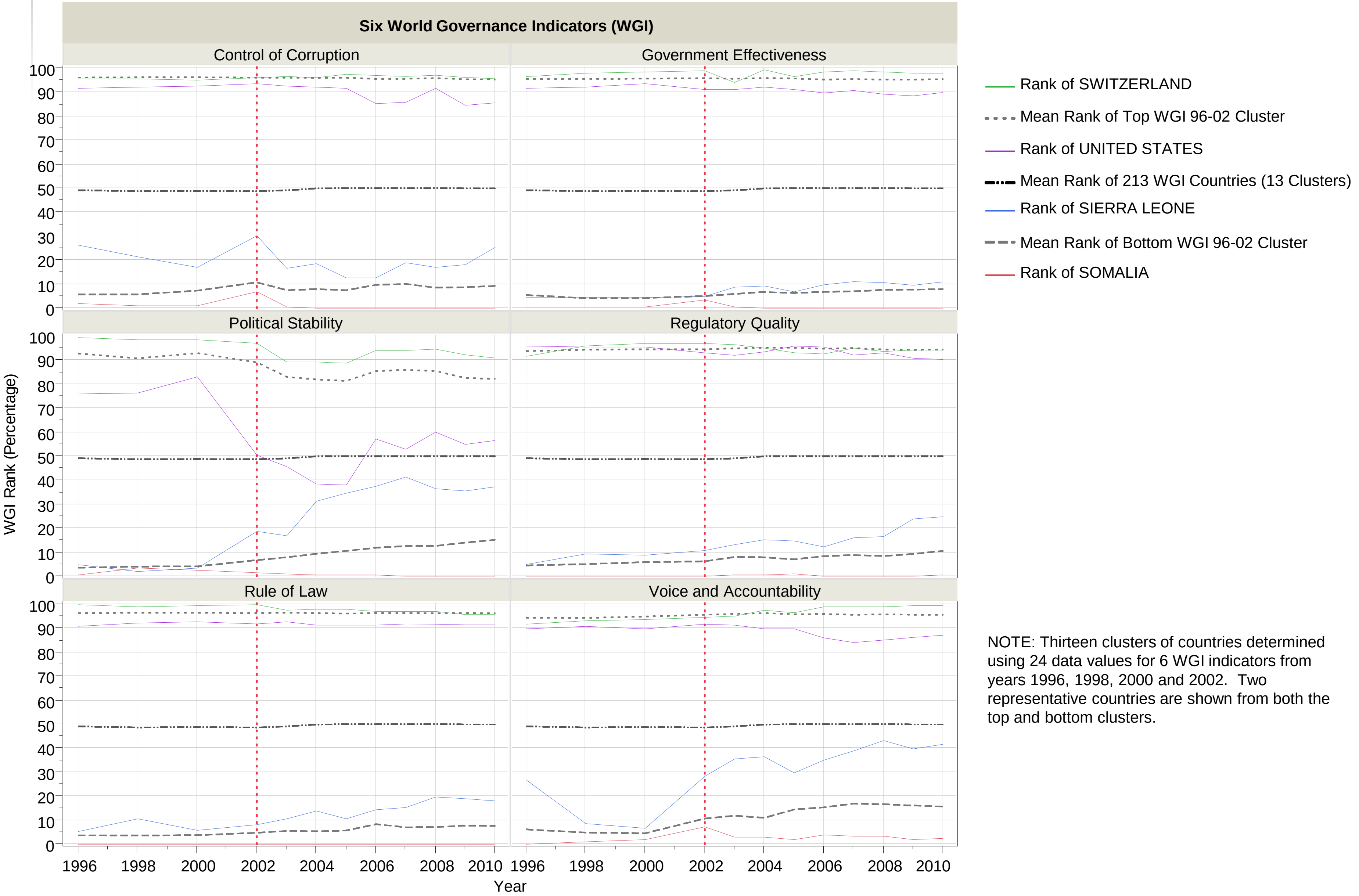
1. Voice and Accountability (VA)
2. Political Stability (PS) and Absence of Violence/Terrorism
3. Government Effectiveness (GE)
4. Regulatory Quality (RQ)
5. Rule of Law (RL)
6. Control of Corruption (CC)

The analysis was performed for the 213 countries in the Worldwide Governance Indicators, 2011 Update data set. The data set can be downloaded from the following link: www.govindicators.org. These are the six aggregate indicators of broad dimensions of governance:

The 24 columns in the heat map are color coded based on the values of the 6 aggregate indicators (CC, GE, PS, RL, RQ, & VA) for the 4 years 1996, 1998, 2000, and 2002. The 12 lowest scoring countries are grouped in cluster #1 shaded red at the top of the chart. The 17 highest scoring countries are grouped in cluster #13 shaded green at the bottom of the chart.

COMPARING WGI RANK PERCENTAGE FOR 2 PAIRS OF 213 COUNTRIES FROM MOST & LEAST STABLE CLUSTERS

Six WGI Indicator Ranks (%) vs. Year for 4 Countries: United States, Switzerland, Sierra Leone & Somalia
(Shown for reference are Mean Rank of all 213 countries and Mean Ranks of Top and Bottom of 13 Clusters)

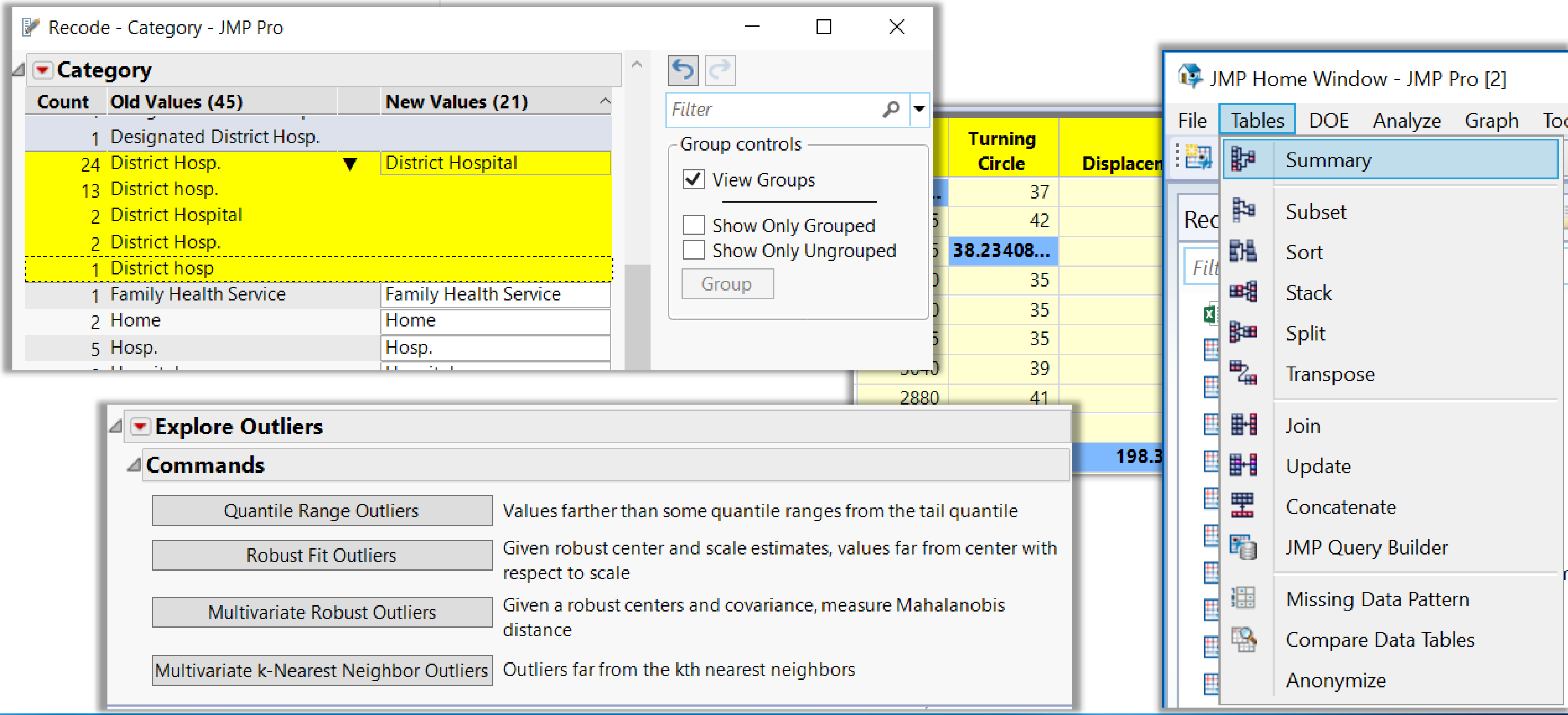


NOTE: Thirteen clusters of countries determined using 24 data values for 6 WGI indicators from years 1996, 1998, 2000 and 2002. Two representative countries are shown from both the top and bottom clusters.

DATA WRANGLING

RECODE, OUTLIER DETECTION, AND IMPUTE MISSING VALUES, STACK, SPLIT, ETC.

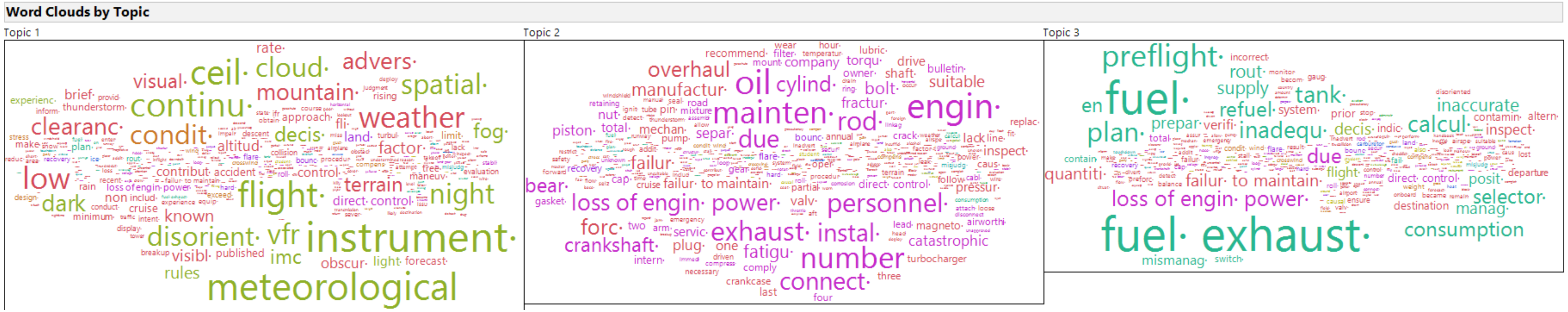
*"60% TO 95% OF THE TIME IS SPENT PREPARING THE DATA"**

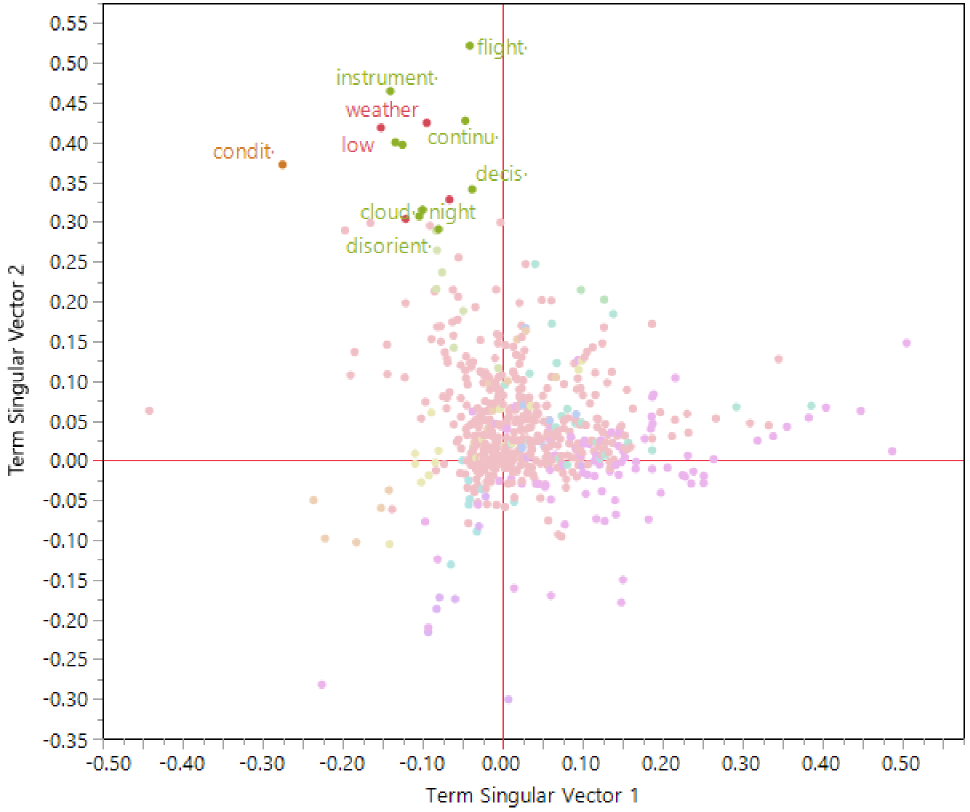


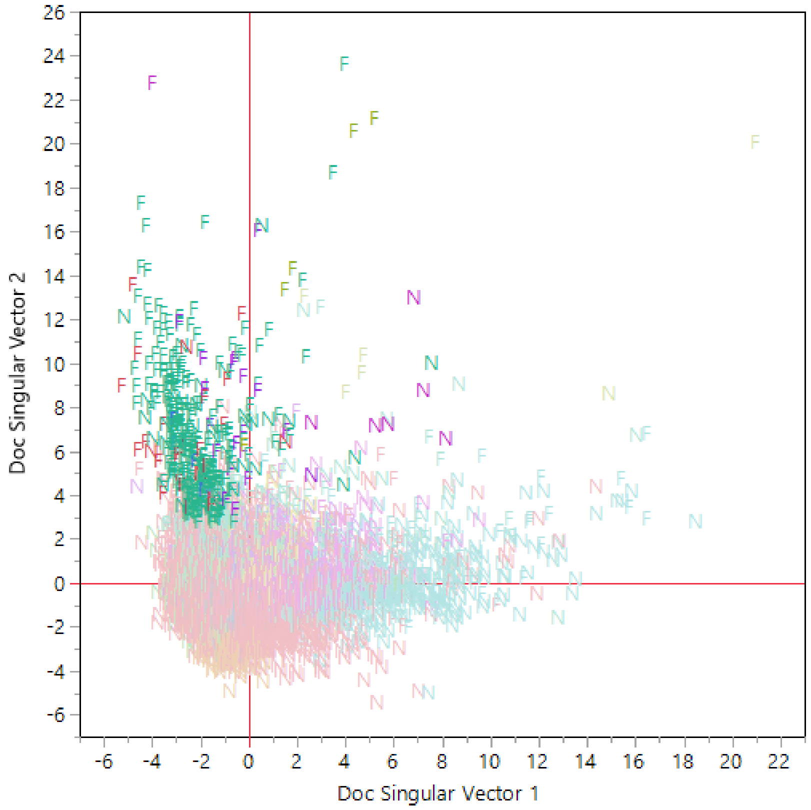
** "85.32% OF ALL STATISTICS ARE MADE UP" – PROF. DICK DE VEAUX*

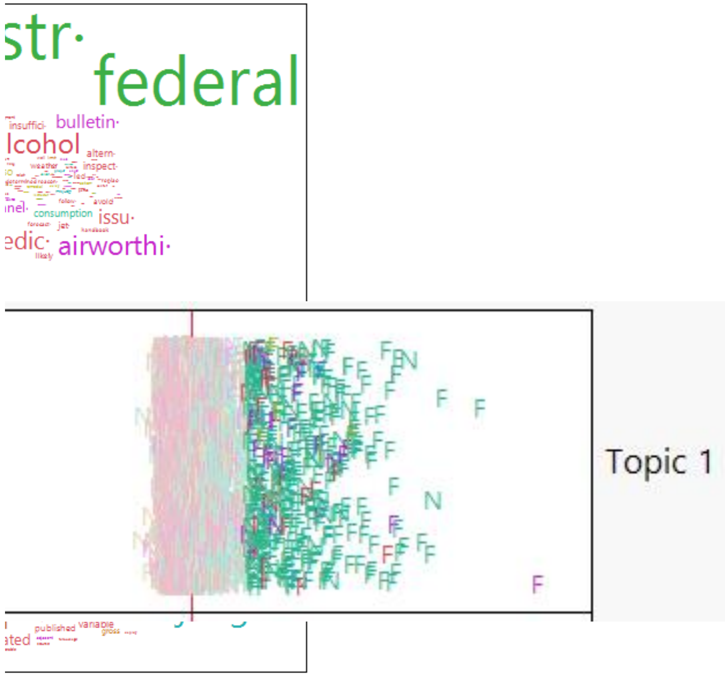
EXPLORATORY TEXT ANALYSIS

DIMENSION REDUCTION OF SPARSE DOCUMENT TERM MATRIX INTO DOCUMENT AND TERM VECTORS – ALSO CLUSTERING OF DOCUMENTS AND TOPICS



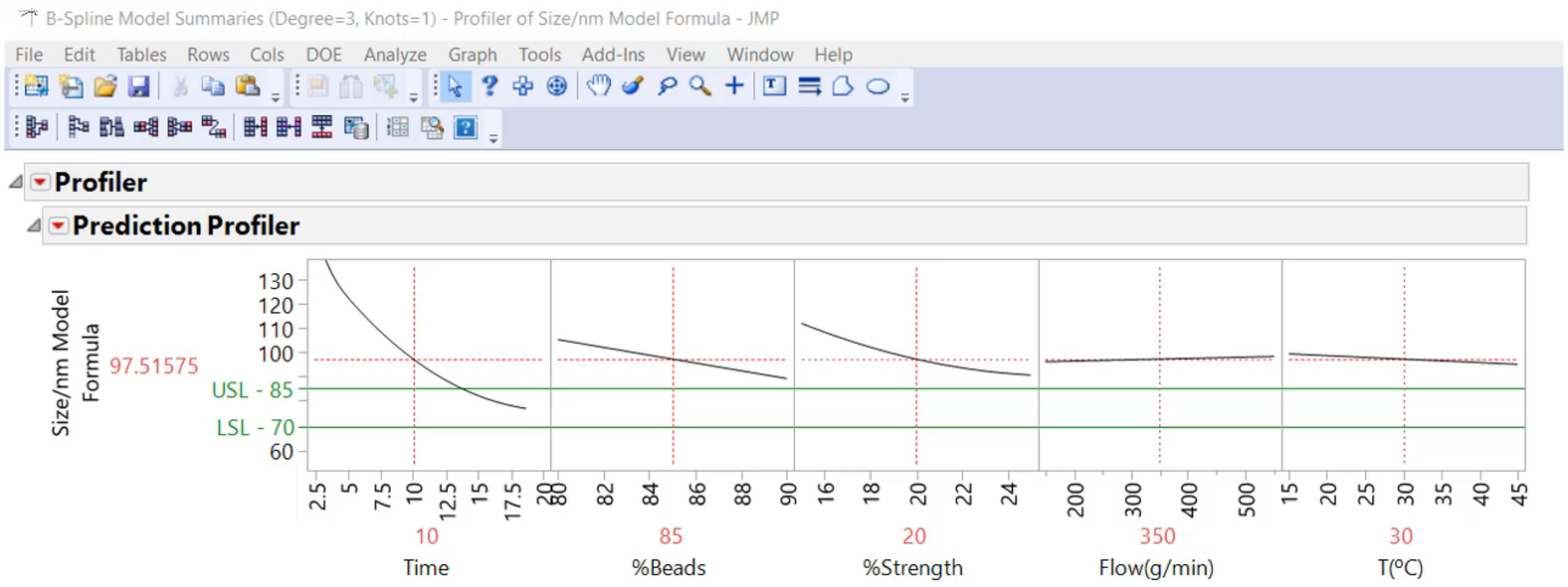






FUNCTIONAL DATA ANALYSIS

MODELING THE "SHAPE" OF A STREAM OF DATA – SHAPE IS THE FUNDAMENTAL UNIT OF OBSERVATION – DIMENSION REDUCTION WITH FUNCTIONAL PCA ABLE TO CONTROL AND PREDICT SHAPE AS FUNCTION OF DOE FACTORS



Functional Data ***Curve***

DoE Factors

THREE TAKEAWAYS

I. Don’t just model for *Best **Prediction***, also seek the *Most **Understanding***

II. ***Prevent Overfitting*** Models Using Training, Validation, and Test Subsets

III. Robust 3-Step Machine Learning Strategy

   1. Use Bootstrap (Random) Forest to avoid missing a variable
   2. Use other Machine Learning methods to create Best Prediction Model (often a neural net is most flexible, but not always)
   3. Use Penalized Regression methods (e.g. LASSO) to get a more interpretable model – sacrifice some accuracy for improved understanding

JMP Defense & Aerospace Team

Anna-Christina De La Iglesia
JMP Program Manager
anna-christina.delaiglesia@jmp.com
919-531-2593

Procurement, Upgrades, License Renewals…

Sam Tobin
JMP Senior Account Representative
sam.tobin@jmp.com
919-531-0640

Technical Questions, Getting Started, Tutorials, Mentoring…

Tom Donnelly, PhD, CAP
JMP Principal Systems Engineer & Co-Insurrectionist
tom.donnelly@jmp.com
302-489-9291