Strategies for Building Predictive Models

Dean Abbott
Abbott Analytics, Inc.
KNIME User Group Meeting
February 14, 2014

Email: dean@abbottanalytics.com
Blog: http://abbottanalytics.blogspot.com
Twitter: @deanabb

Instructor - Dean Abbott

- Education
  - Master of Applied Mathematics, University of Virginia
  - B.S. Computational Mathematics, Rensselaer Polytechnic Institute
- Applied Data Mining for 25+ years in
  - Tax Compliance, Fraud Detection
  - Text Mining and Concept Classification
  - Direct Marketing, CRM, Survey Analysis, Market Basket Analysis
  - Predictive Toxicology, Biological Risk Assessment
  - Earlier
    - Signal and Image Processing, Guidance and Control
    - Optical Character Recognition & Postnet Bar Code Readers
- Data Mining Course Instruction
  - Taught dozens of short courses, conference tutorials, lectures, in-house custom courses

Strategies

1. Know what we are doing
2. Have a plan for the project
3. Assess models the way you want to use them
4. Do for the algorithms what they cannot do for themselves
5. Deploy models wisely

Strategy 1:
Know What You are Doing

- What is Predictive Analytics?
- How does PA differ from
  - Statistics
  - BI
  - Big Data

What is Predictive Analytics?

- Wikipedia Definitions
  - Predictive analytics is an area of statistical analysis that deals with extracting information from data and uses it to predict future trends and behavior patterns.
  - The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting it to predict future outcomes.

What is Predictive Analytics?

- Other Definitions (in the news and blogs)
  - Predictive Analytics is emerging as a game-changer. Instead of looking backward to analyze "what happened?" predictive analytics help executives answer "What's next?" and "What should we do about it?" (Forbes Magazine, April 1, 2010)
  - Predictive analytics is the branch of data mining concerned with the prediction of future probabilities and trends. (searchcrm.com)
  - Predictive Analytics *is* data mining re-badged because too many people were claiming to do data mining and weren't. (Tim Manns paraphrasing Wayne Erickson of TDWI)

What is Predictive Analytics?
Simple Definitions

- *Data driven* analysis for [large] data sets
  - Data-driven to discover input combinations
  - Data-driven to validate models
- *Automated* pattern discovery
  - Key input variables
  - Key input combinations

Statistics vs. Predictive Analytics

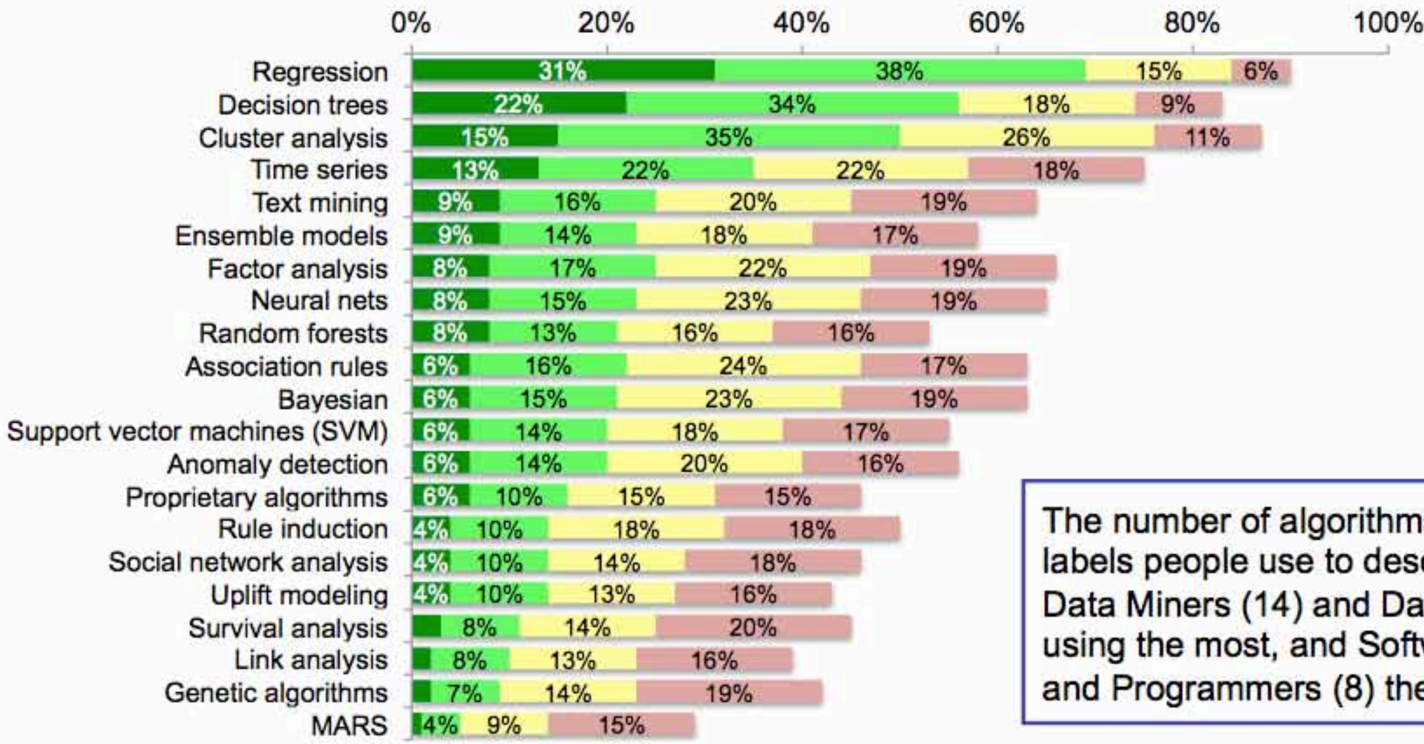
	Statistics	Predictive Analytics
View of the "other" field	"data dredging"	"we can do *that* … and more!"
Emphasis	Theory; Optimum Solutions	"Good" Heuristics
Approach	Parametric	Non-parametric
Key Metrics of Performance	R^2, p-values, S.E.	Lift, ROC
What is King?	Model	Data

See David J. Hand, "Statistics and Data Mining: Intersecting Disciplines", SIGKDD Explorations, Vol. 1, No. 1, June 1999, pp. 16-19.

Business Intelligence vs. Predictive Analytics

	Business Intelligence	Predictive Analytics
View of the "other" field	"we're the foundation, (they're so complicated!)"	"they report the past, we predict the future!"
Emphasis	What happened?	What do we think *will* happen?
Approach	User-driven Reporting	Algorithms, Searching
Key Metrics of Performance	KPIs	Lift, ROC
What is King?	Data (via Analyst)	Data (via Algorithms)

Rexer Analytics Survey (2013): Predictive Analytics Algorithms



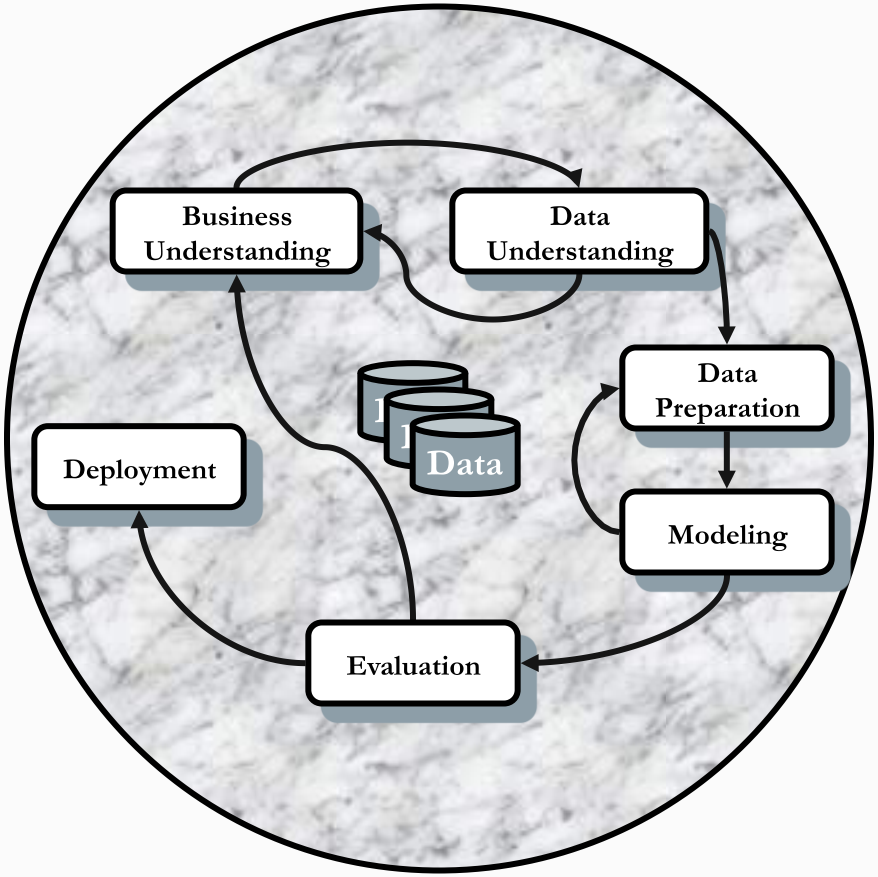
The number of algorithms used varies by the labels people use to describe themselves, with Data Miners (14) and Data Scientists (14) using the most, and Software Developers (9) and Programmers (8) the fewest.

Strategy 2:
Have a Plan

- Use CRISP-DM
  - Or similar framework
- Don't be too strict
  - These are suggested steps, not recipes

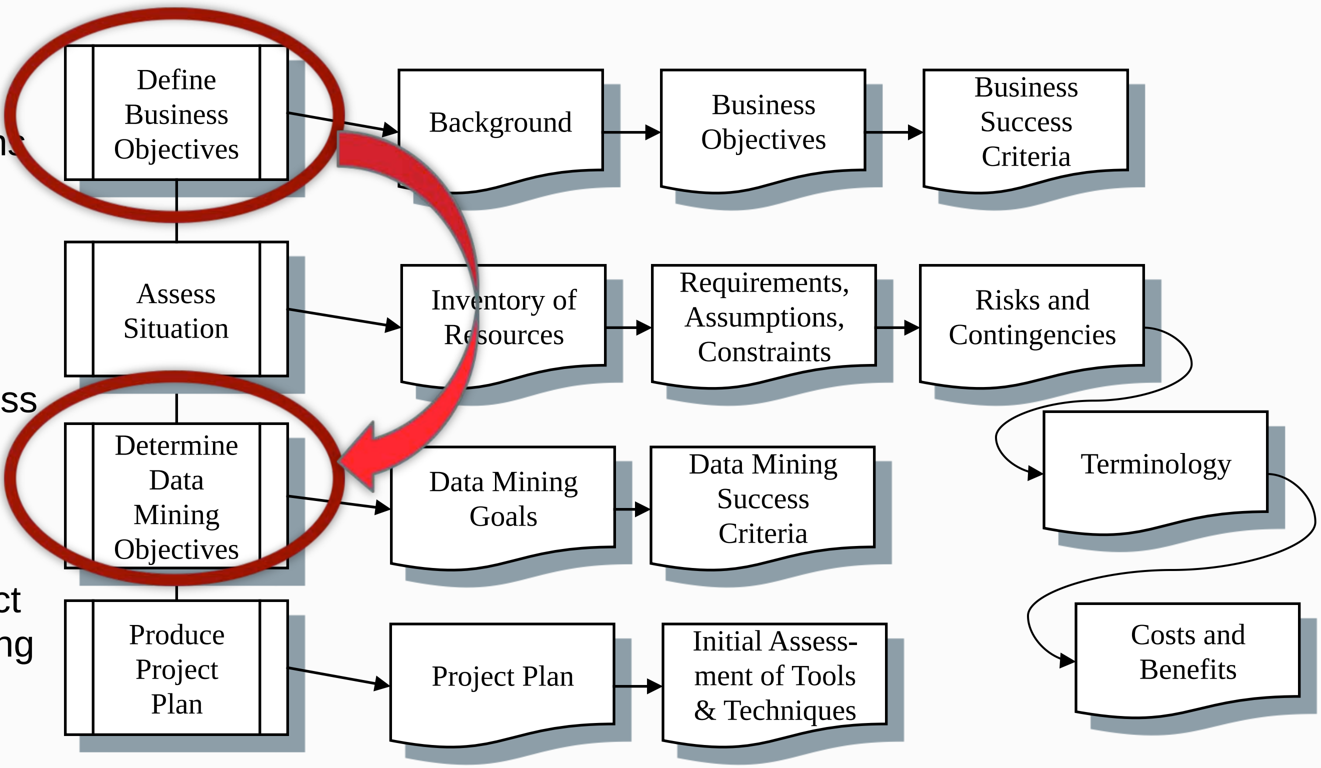
What do Predictive Modelers do? The CRISP-DM Process Model

- CRoss-Industry Standard Process Model for Data Mining
- Describes Components of Complete Data Mining Cycle from the Project Manager's Perspective
- Shows Iterative Nature of Data Mining

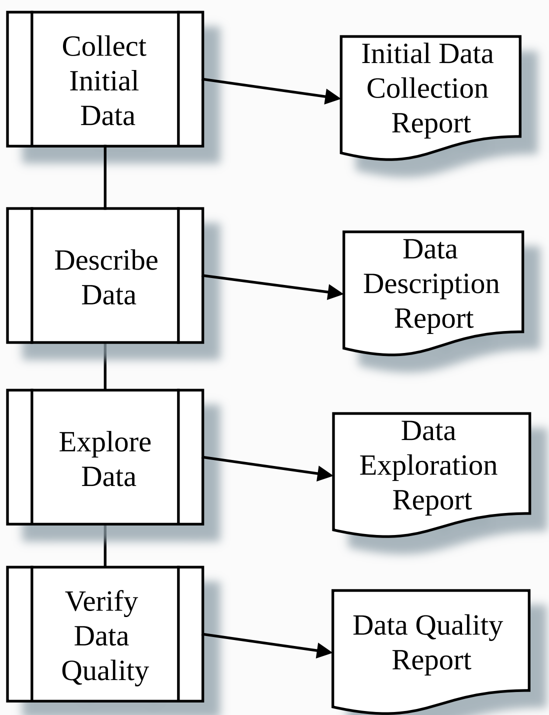


CRISP-DM: Business Understanding Steps

- Ask Relevant Business Questions
- Determine Data Requirements to Answer Business Question
- Translate Business Question into Appropriate Data Mining Approach
- Determine Project Plan for Data Mining Approach



CRISP-DM Step 2:
Data Understanding Steps

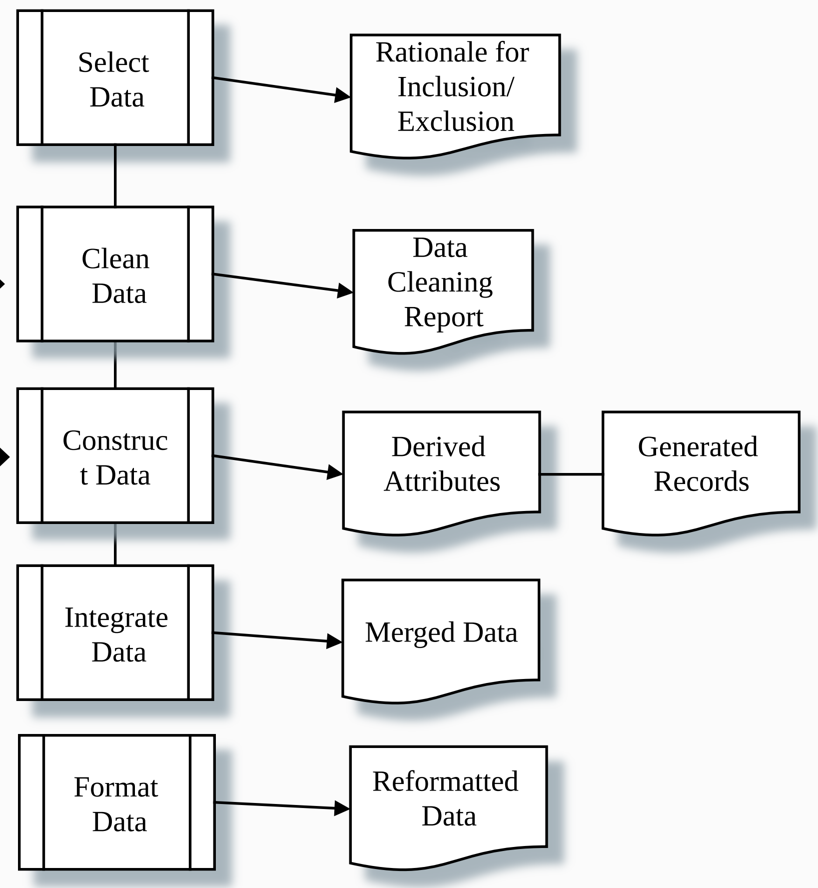


- Collect initial data
  - Internal data: historical customer behavior, results from previous experiments
  - External data: demographics & census, other studies and government research
  - Extract superset of data (rows and columns) to be used in modeling
  - Identify form of data repository: multiple vs. single table, flat file vs. database, local copy vs. data mart
- Perform Preliminary Analysis
  - Characterize Data (describe, explore, verify)
  - Condition Data

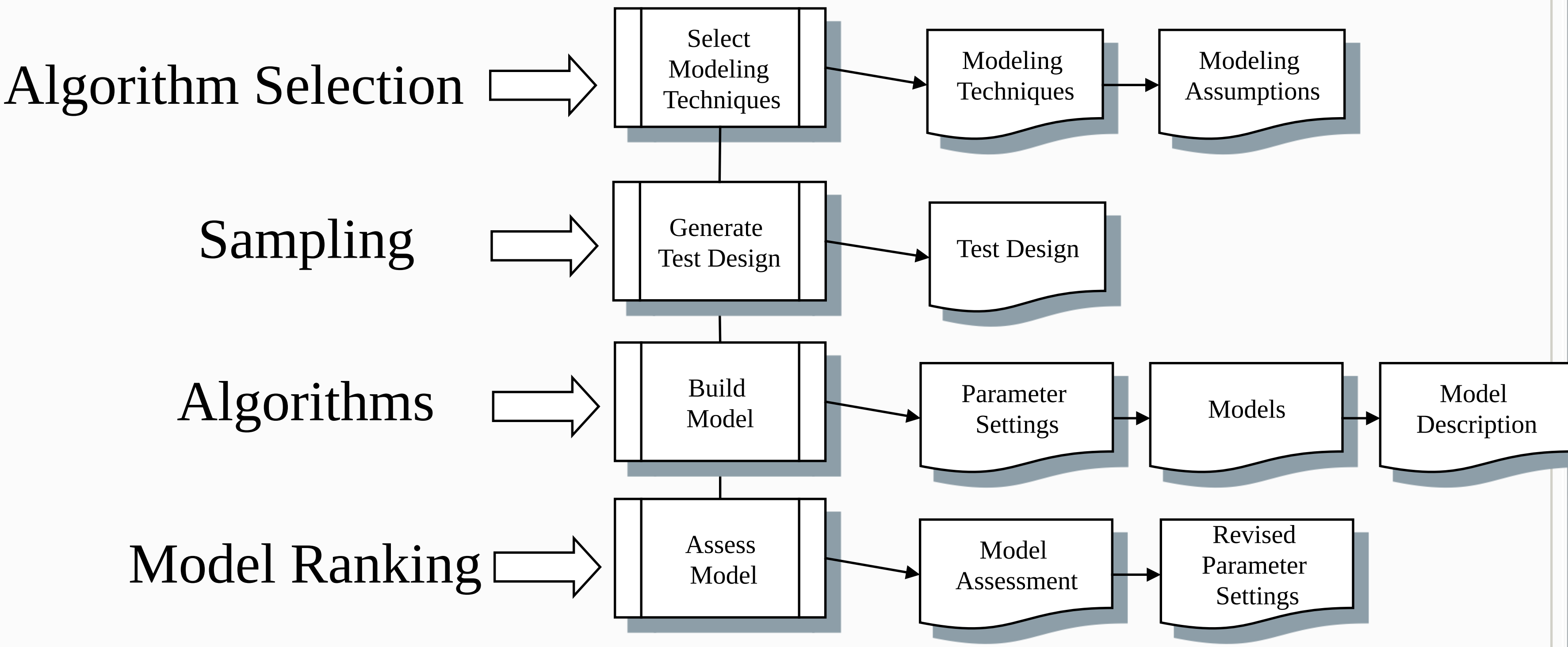
CRISP-DM Step 3:
Data Preparation (Conditioning) Steps

Fix Data Problems

Create Features



CRISP-DM Step 4:
Modeling Steps



CRISP-DM vs. SEMMA

- CRISP-DM
  Six Steps
- SEMMA
  Five Steps

Table 2.3. Comparison of methods

CRISP	SEMMA	Nayak & Qiu
Business understanding	Assumes well-defined question	Goals were defined Develop tools to better utilize problem reports
Data understanding	*Sample* *Explore*	Looked at data in problem reports
Data preparation	*Modify data*	Data pre-processing Data cleaning Data transformation
Modeling	*Model*	Data modeling
Evaluation	*Assess*	Analyzing results
Deployment		

Table from Advanced Data Mining Techniques, Olsen and Delen, Springer, 2008

Strategy 3: Assess Models the Way you Use Them

- Standard Assessment Methods
  - Batch Methods
  - Rank-Ordered Methods
- Why the Method Matters
  - Outliers
  - Sampling and Accuracy

Classification Accuracy from Decision Thresholds

- If P(Target_B = 1) is greater than a pre-defined threshold, the prediction is Target_B = 1.
- If the prediction matches the actual Target_B value, the decision is "correct". Otherwise it is wrong
- With the threshold of 0.05,
  - first 17 records are above the threshold
  - 9 records have "correct" predictions
  - 8 records have "incorrect" predictions

P(Target_B = 1)	CONTROLN	LastGift	TARGET_B	P(Target_B = 0)
0.0731	185436	0	0	0.9269
0.0715	14279	1	1	0.9285
0.0699	727	2	1	0.9301
0.0683	24610	3	1	0.9317
0.0668	22645	4	1	0.9332
0.0653	82943	5	1	0.9347
0.0639	108412	6	0	0.9361
0.0624	190313	7	1	0.9376
0.0611	48804	8	0	0.9389
0.0597	123822	9	1	0.9403
0.0583	94039	10	0	0.9417
0.0570	47605	11	0	0.9430
0.0558	25641	12	1	0.9442
0.0545	47476	13	0	0.9455
0.0533	6023	14	0	0.9467
0.0521	47784	15	0	0.9479
0.0509	148569	16	1	0.9491
0.0497	171099	17	0	0.9503

If (P(Target_B = 1) > 0.05
Then 1
Else 0

Typical Binary Classification Accuracy Metrics

Confusion Matrix		Predicted Class		
		0 (predicted value is negative)	**1** (predicted value is positive)	Total Actual (down)
Actual Class	**0** (actual value is negative)	$\mathbf{t_n}$ (true negative)	$\mathbf{f_n}$ (false negative, false dismissal)	Total actual negatives tn + fn
	1 (actual value is positive)	$\mathbf{f_p}$ (false positive, false alarm)	$\mathbf{t_p}$ (true positive)	Total actual positives tp + fp
	Total Predicted (across)	Total negative predictions tn + fp	Total positive predictions tp + fn	Total Examples tp + tn + fp + fn

$$PCC = \frac{t_p + t_n}{t_p + t_n + f_p + f_n}$$

$$Precision = \frac{t_p}{t_p + f_p}$$

$$Recall = \frac{t_p}{t_p + f_n}$$

$$False\ Alarm\ Rate\ (FA) = \frac{f_p}{t_n + f_p}$$

$$False\ Dismissal\ Rate\ (FD) = 1 - Recall = \frac{f_n}{t_p + f_n}$$

$$Sensitivity = Recall = \frac{t_p}{t_p + f_n}$$

$$Specificity = True\ Negative\ Rate = \frac{t_n}{t_n + f_p}$$

$$Type\ I\ Error = \frac{f_p}{t_p + t_n + f_p + f_n}$$

$$Type\ II\ Error = \frac{f_n}{t_p + t_n + f_p + f_n}$$

Percent Correct Classification (PCC)

$$PCC = (t\_n + t\_p)/(t\_p + t\_n + f\_p + f\_n)$$

Confusion Matrix		Predicted Class		
		0 (predicted value is negative)	**1** (predicted value is positive)	Total Actual (down)
Actual Class	**0** (actual value is negative)	t_n (true negative)	f_p (false positive, false alarm)	Total actual negatives tn + fp
	1 (actual value is positive)	f_n (false negative, false dismissal)	t_p (true positive)	Total actual positives tp + fn
	Total Predicted (across)	Total negative predictions tn + fn	Total positive predictions tp + fp	Total Examples tp + tn + fp + fn

Batch vs. Rank-Ordered Approaches to Model Evaluation

- "Batch" approaches
  - Score every record with the same weight
  - Provide a summary for the entire scored data set
    - PCC, Precision, Recall, Type I Errors, Type II Errors, etc.

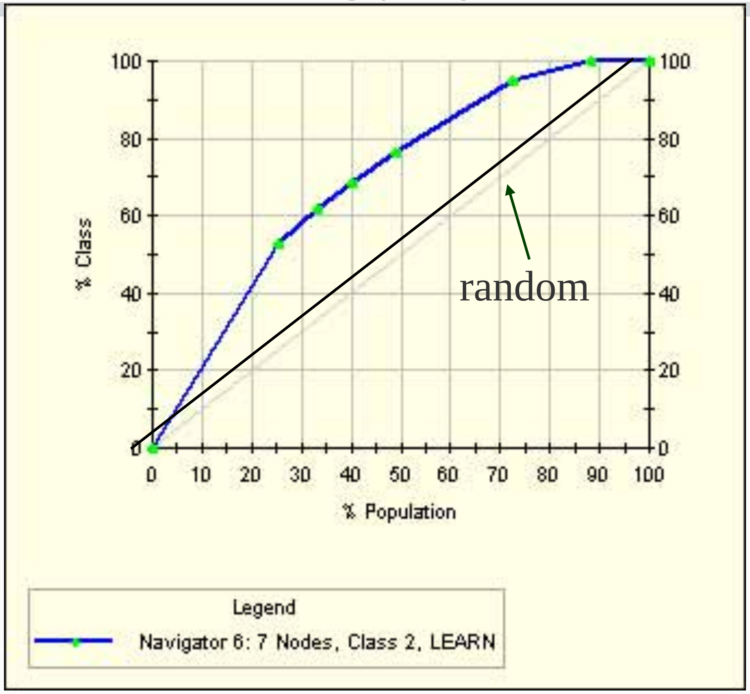
- Rank-ordered approaches
  - Sort population by model scores, high to low
  - Accumulate a metric as one goes down the ordered file
  - Reports results by group, typically deciles, demi-deciles, percentiles, etc.
  - Examples
    - Lift, Gains, ROC, Profit

Three Key Rank-Ordered Metrics

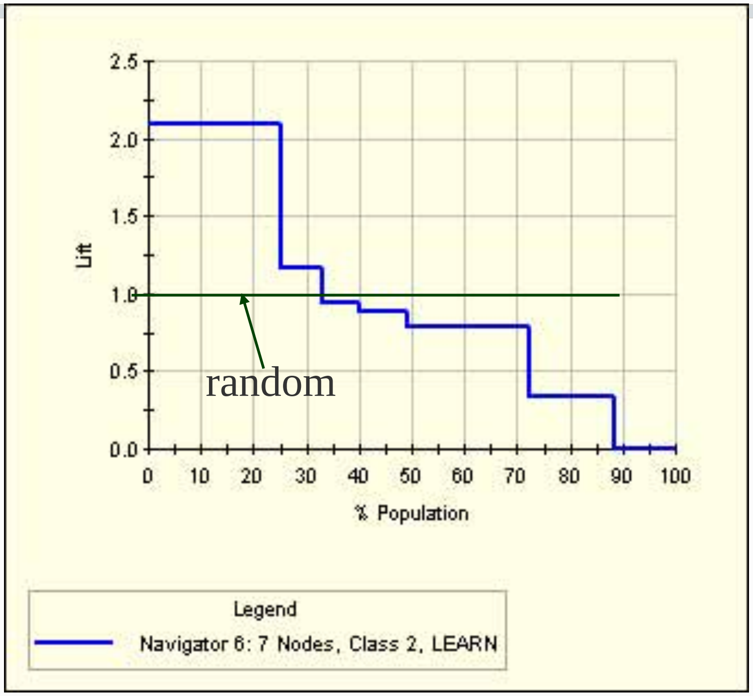
Measure	What is Measures
Gain	% of target records found
Lift	ratio of target gain to average response rate
ROC	change classifier probability threshold from 0 to 1; sensitivity vs. 1 - specificity for each threshold

Gains Charts and Lift Curves

Gains



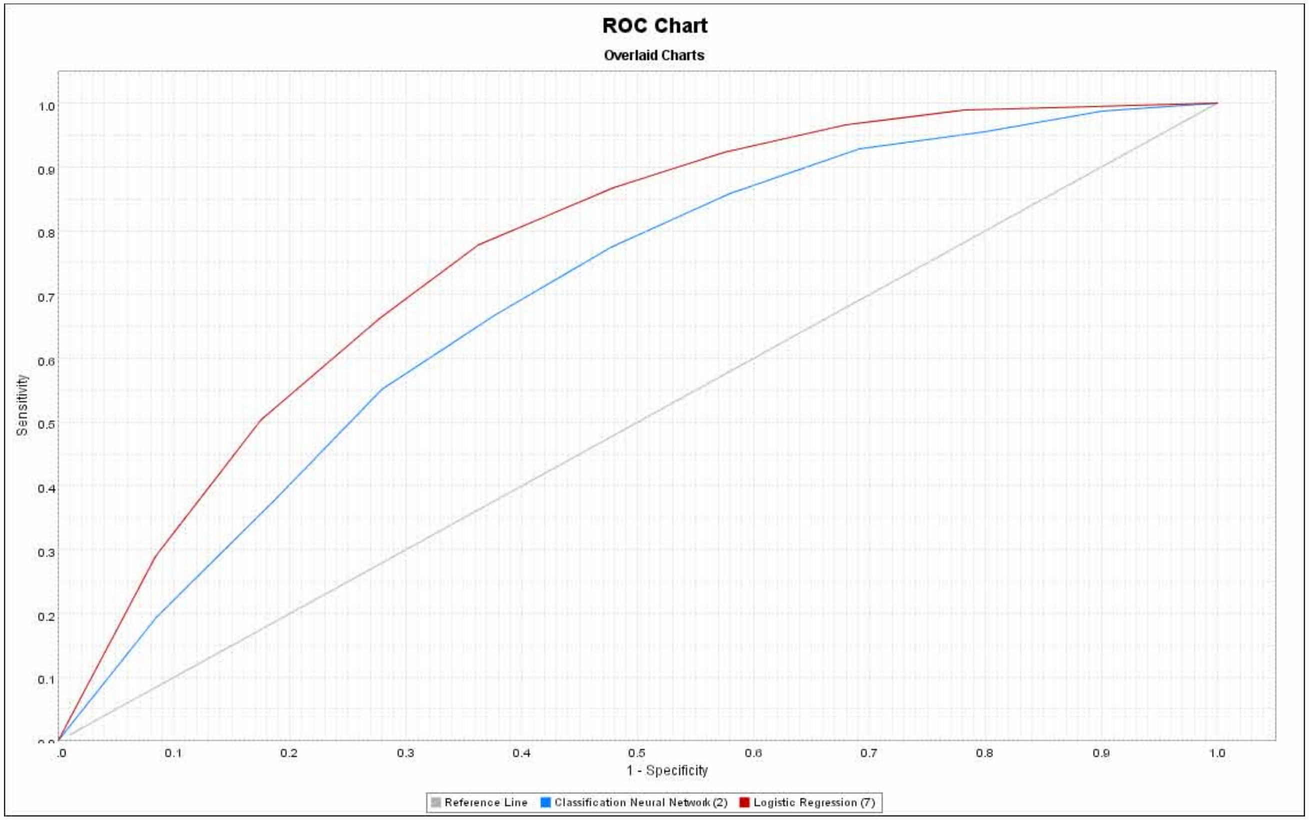
Lift



- Number Respondants are Rank-ordered (sorted) by predicted values
- X axis is the percentage of records as go down file.
- Gain is the pct. of target=1 found at indicated file depth
- Lift ratio is how many times more respondants at given customer depth compared to random selection
- Random gain has slope equal to proportion of respondants in training data
- Random lift is 1.0

ROC Curves

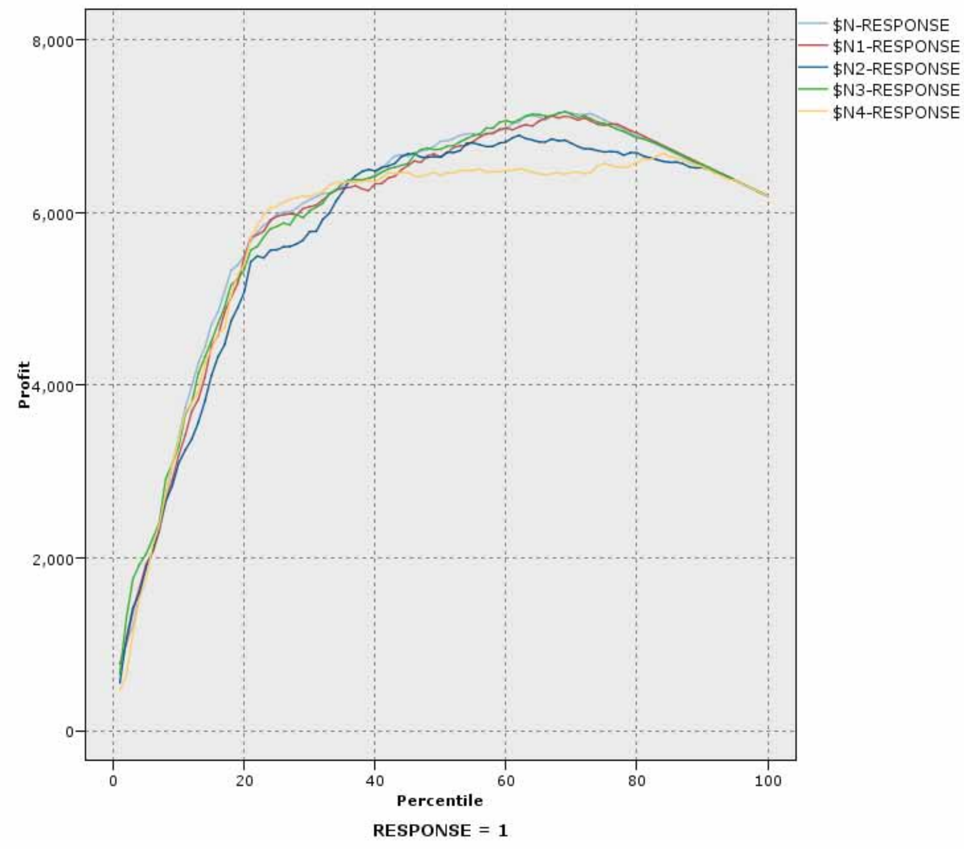
Sensitivity (True Alerts)



1 – Specificity (False Alarms)

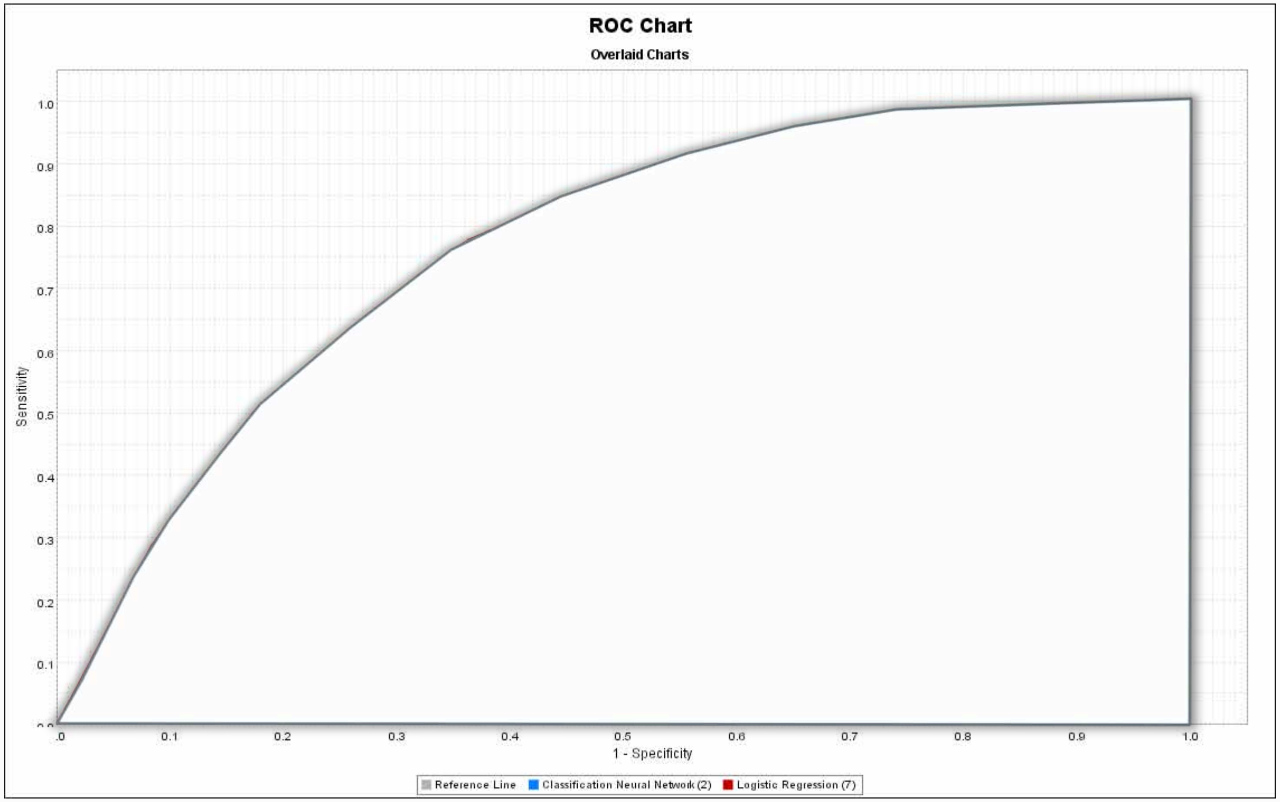
Profit / ROI Charts

- Profit
  - = Revenue - Cost
  - Cumulative Measure
  - Increases as long as ROI > 0 in segments
  - Negative => campaign loses money
- ROI Percentage
  - = Profit / Cost
    - or (Revenue / Cost) -1
  - Cumulative Measure
  - Negative => campaign loses money



26

Area Under the (ROC) Curve (AUC)



- AUC = 1.0 is a perfect model
- AUC = 0.5 is a random model
- Gini Index = 2*AUC – 1
  or
  AUC = (Gini Index + 1) / 2

$$G_1 = 2AUC - 1$$

27

The Conflict with Predictive Modeling Algorithm Objectives

Algorithm Objectives

- Linear Regression and Neural networks minimize squared error
- C5 maximizes Information Gain
- CART maximizes Gini index
- Logistic regression maximizes the log of the odds of the probability the record belongs to class "1" (classification accuracy)
- Nearest neighbor minimizes Euclidean distance

Sample Business Objectives

- Maximize net revenue
- Contact max # customers to achieve response rate of 13%
- Maximize responders subject to a budget of $100,000
- Maximize recovered revenue from customers likely to churn
- Maximize collected revenue by identifying next best case to collect
- Minimize false alarms in 100 transactions most likely to be fraudulent

PAKDD Cup 2007 Results: Score Metric Changes Winner

- Correlation of AUC rank with top Decile Rank: 0.76

Modeling Technique ->	Modeling Implementation ->	Participant Affiliation Location	Participant Affiliation Type	AUCROC (Trapezoidal Rule)	AUCROC (Trapezoidal Rule) Rank	Top Decile Response Rat	Top Decile Response Rate Rank
TreeNet + Logistic Regression	Salford Systems	Mainland China	Practitioner	70.01%	1	13.00%	7
Probit Regression	SAS	USA	Practitioner	69.99%	2	13.13%	6
MLP + n-Tuple Classifier		Brazil	Practitioner	69.62%	3	13.88%	1
TreeNet	Salford Systems	USA	Practitioner	69.61%	4	13.25%	4
TreeNet	Salford Systems	Mainland China	Practitioner	69.42%	5	13.50%	2
Ridge Regression	Rank	Belgium	Practitioner	69.28%	6	12.88%	9
2-Layer Linear Regression		USA	Practitioner	69.14%	7	12.88%	9
Log Regr+ Decision Stump + AdaBoost + VFI		Mainland China	Academia	69.10%	8	13.25%	4
Logistic Average of Single Decision Functions		Australia	Practitioner	68.85%	9	12.13%	17
Logistic Regression	Weka	Singapore	Academia	68.69%	10	12.38%	16
Logistic Regression		Mainland China	Practitioner	68.58%	11	12.88%	9
Decision Tree + Neural Network + Logistic Regression		Singapore		68.54%	12	13.00%	7
Scorecard Linear Additive Model	Xeno	USA	Practitioner	68.28%	13	11.75%	20
Random Forest	Weka	USA		68.04%	14	12.50%	14
Expanding Regression Tree + RankBoost + Bagging	Weka	Mainland China	Academia	68.02%	15	12.50%	14
Logistic Regression	SAS + Salford	India	Practitioner	67.58%	16	12.00%	19
J48 + BayesNet	Weka	Mainland China	Academia	67.56%	17	11.63%	21
Neural Network + General Additive Model	Tiberius	USA	Practitioner	67.54%	18	11.63%	21
Decision Tree + Neural Network		Mainland China	Academia	67.50%	19	12.88%	9
Decision Tree + Neural Network + Log. Regression	SAS	USA	Academia	66.71%	20	13.50%	2

http://lamda.nju.edu.cn/conf/pakdd07/dmc07/results.htm

Model Comparison:
Different Metrics Tell Different Stories

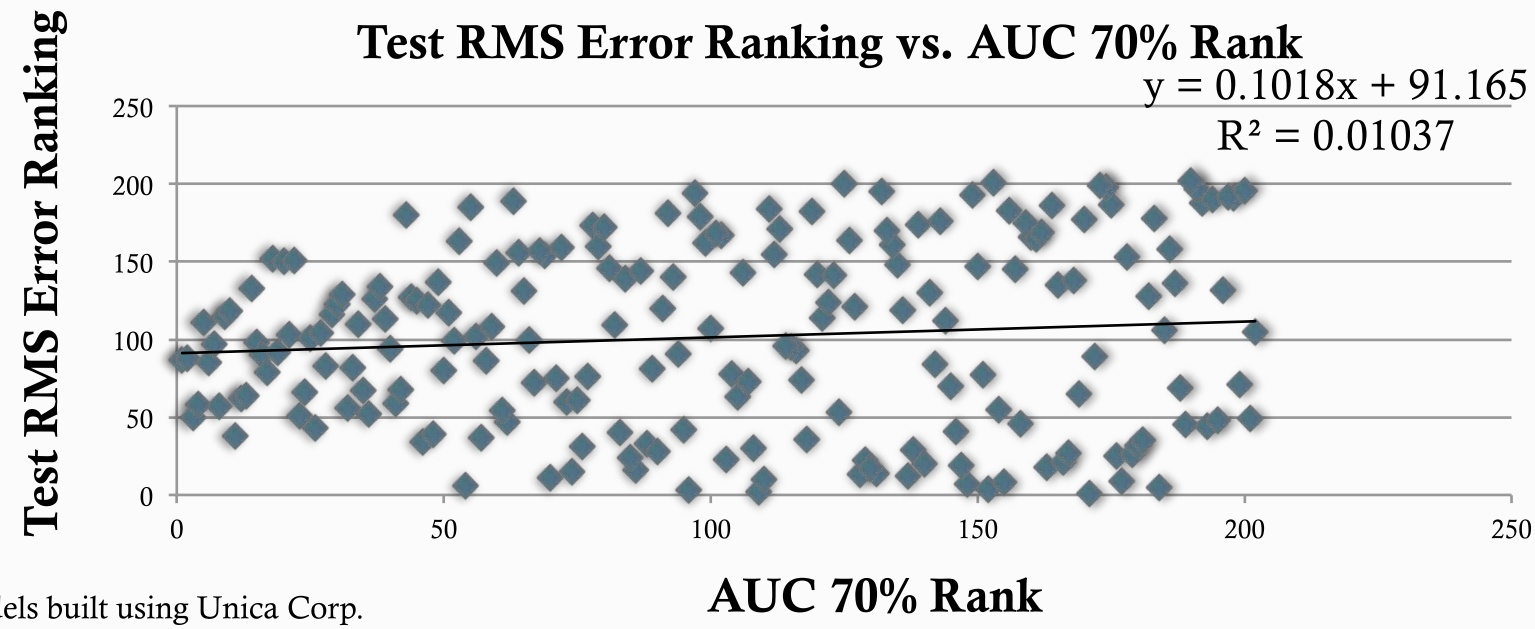
Model Number	Model ID	AUC	Train RMS	Test RMS	AUC Rank	Train RMS Rank	Test RMS Rank
50	NeuralNet1032	73.3%	0.459	0.370	9	53	1
39	NeuralNet303	72.4%	0.477	0.374	42	59	2
36	NeuralNet284	75.0%	0.458	0.376	2	52	3
31	NeuralNet244	72.7%	0.454	0.386	33	49	4
57	CVLinReg2087	70.4%	0.397	0.393	52	5	5
34	NeuralNet277	72.7%	0.455	0.399	28	50	6
37	NeuralNet297	72.4%	0.449	0.399	43	38	7
56	CV_CART2079	68.0%	0.391	0.401	54	4	8
54	CVNeuralNet2073	67.9%	0.403	0.401	55	6	9
59	CVNeuralNet2097	66.0%	0.403	0.401	59	7	10
61	CV_CART2104	70.4%	0.386	0.402	53	3	11
42	NeuralNet334	72.4%	0.450	0.404	40	44	12
52	CVLinReg2063	67.5%	0.404	0.404	57	8	13
41	NeuralNet330	72.4%	0.443	0.406	41	16	14
38	NeuralNet300	72.4%	0.451	0.408	38	45	15
55	CV_CHAID2078	64.6%	0.380	0.411	60	2	16
45	NeuralNet852	74.2%	0.456	0.413	3	51	17
53	CVLogit2068	67.5%	0.414	0.414	58	10	18
60	CV_CHAID2102	61.5%	0.380	0.414	61	1	19
58	CVLogit2092	67.7%	0.413	0.414	56	9	20

- Top RMS model is 9th in AUC, 2nd Test RMS rank is 42nd in AUC
- Correlation between rankings:

	AUC Rank	Train RMS Rank	Test RMS Rank
AUC Rank	1		
Train RMS Rank	(0.465)	1	
Test RMS Rank	(0.301)	0.267	1

KDDCup 98 Data: Top 200 Models Built Using Stepwise Variable Selection

- Error Metrics
  - Root-Mean-Squared (RMS) Error on Test data
  - Area Under the Curve (AUC) at the 70% Depth



Models built using Unica Corp. Affinium Model

How Sampling Effects Accuracy Measures

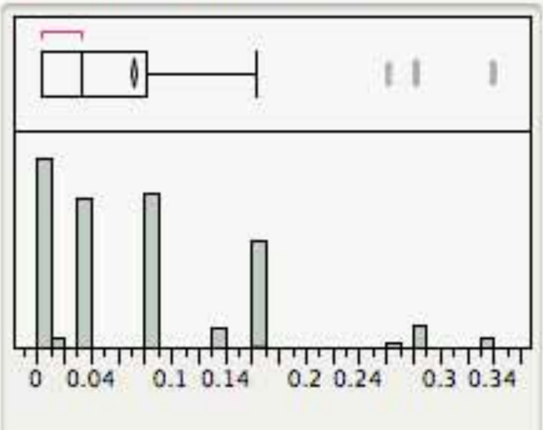
- For example, 95% non-responders (N), 5% responders (R)
- What's the Problem? (The justification for resampling)
  - "Sample is biased toward responders"
  - "Models will learn non-responders better"
  - "Most algorithms will generate models that say 'call everything a non-responder' and get 93% correct classification!" (I used to say this too)
- Most common solution:
  - Stratify the sample to get 50%/50% (some will argue that one only needs 20-30% responders)

What the Predictions Looks Like

Decision Tree

Distributions

Prob(RESPONSE_STR==R)



Quantiles

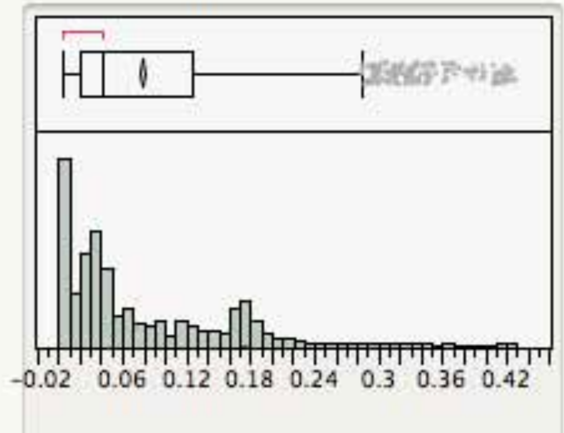
Summary Statistics

Mean	0.0715519
Std Dev	0.0747384
Std Err Mean	0.0010182
Upper 95% Mean	0.0735479
Lower 95% Mean	0.0695558
N	5388

Neural Network

Distributions

Probability(RESPONSE_STR=R)



Quantiles

Summary Statistics

Mean	0.0778526
Std Dev	0.0812571
Std Err Mean	0.001107
Upper 95% Mean	0.0800228
Lower 95% Mean	0.0756824
N	5388

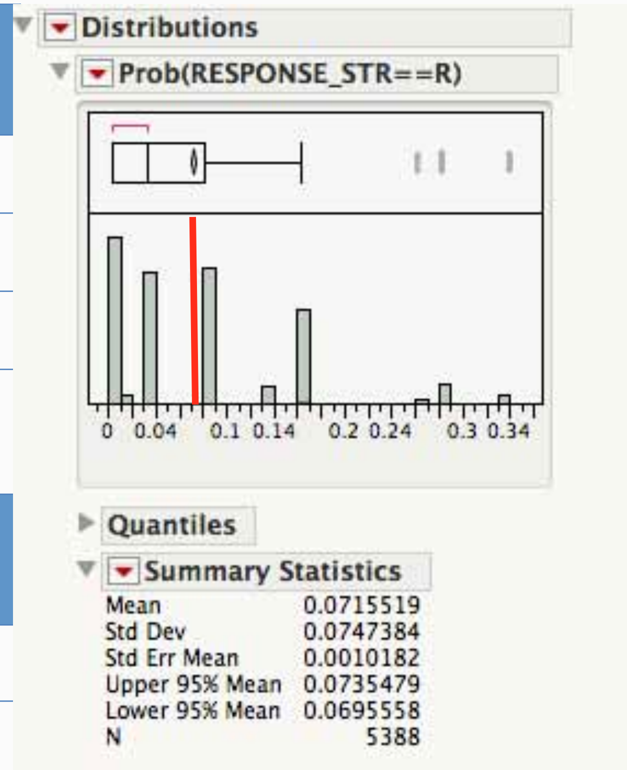
Confusion Matrices For the Decision Tree: Before and After

Decision Tree:
Threshold at 0.5

Response_STR	N	R	Total
N	5,002	0	5,002
R	386	0	386
Total	5,388	0	5,388

Decision Tree:
Threshold at 0.071

Response_STR	N	R	Total
N	2,798	2,204	5,002
R	45	341	386
Total	2,843	2,545	5,388



To KNIME

KNIME Sampling

The Winner is… Best Accuracy

http://www.netflixprize.com/leaderboard

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace_	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59

Why Model Accuracy is Not Enough: Netflix Prize

NETFLIX

http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html

A year into the competition, the Korbell team won the first Progress Prize with an 8.43% improvement. They reported more than 2000 hours of work in order to come up with the final combination of 107 algorithms that gave them this prize. And, they gave us the source code. We looked at the two underlying algorithms with the best performance in the ensemble: *Matrix Factorization* (which the community generally called SVD, *Singular Value Decomposition*) and *Restricted Boltzmann Machines* (RBM). SVD by itself provided a 0.8914 RMSE, while RBM alone provided a competitive but slightly worse 0.8990 RMSE. A linear blend of these two reduced the error to 0.88. To put these algorithms to use, we had to work to overcome some limitations, for instance that they were built to handle 100 million ratings, instead of the more than 5 billion that we have, and that they were not built to adapt as members added more ratings. But once we overcame those challenges, we put the two

If you followed the Prize competition, you might be wondering what happened with the final Grand Prize ensemble that won the $1M two years later. This is a truly impressive compilation and culmination of years of work, blending hundreds of predictive models to finally cross the finish line. We evaluated some of the new methods offline but the additional accuracy gains that we measured did not seem to justify the engineering effort needed to bring them into a production environment. Also, our focus on improving Netflix personalization had shifted to the next level by then. In the remainder of this post we will explain how and why it has shifted.

Why Data Science is Not Enough: Netflix Prize

NETFLIX

http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html

Now it is clear that the Netflix Prize objective, accurate prediction of a movie's rating, is just one of the many components of an effective recommendation system that optimizes our members enjoyment. We also need to take into account factors such as context, title popularity, interest, evidence, novelty, diversity, and freshness. Supporting all the different contexts in which we want to make recommendations requires a range of algorithms that are tuned to the needs of those contexts. In the next part of this post, we will talk in more detail about the ranking problem. We will also dive into the data and models that make all the above possible and discuss our approach to innovating in this space.

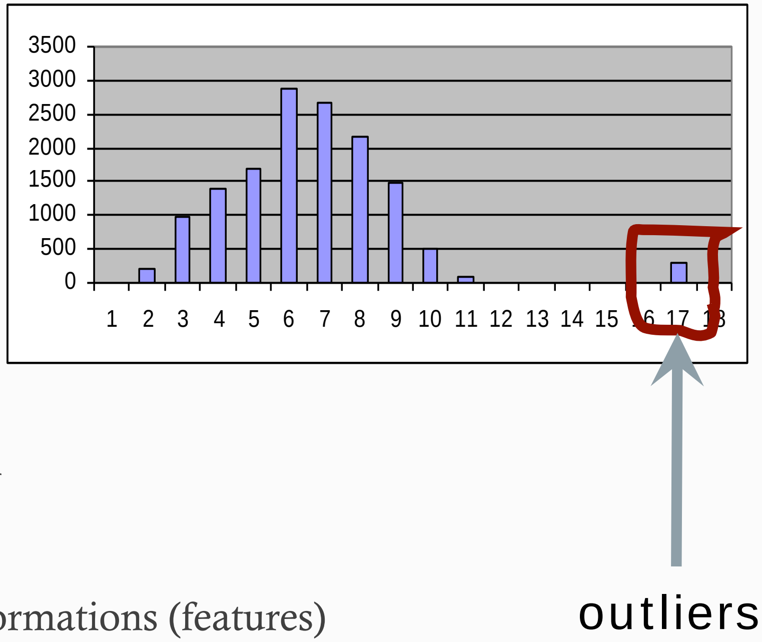
There's more to a solution than accuracy—you have to be able to use it!

Strategy 4
Do for algorithms what they can't do for themselves

- Get the data right
- Understand how algorithms can be fooled with "correct" data
  - Outliers
  - Missing Values
  - Skew
  - High Cardinality

Clean Data: Outliers

- Are the outliers problems?
  - Some algorithms: "yes"
    - Linear regression, nearest neighbor, nearest mean, principal component analysis
    - In other words, algorithms that need mean values and standard deviations
  - Some algorithms: "no"
    - Decision trees, neural networks



- If outliers are problems for the algorithm
  - Are they key data points?
    - *Do not* remove these
    - Consider "taming" outliers with transformations (features)
  - Are they anomalies or otherwise uninteresting to the analysis
    - Remove from data so that they don't bias models

To KNIME

KNIME outlier

Clean Data: Missing Values

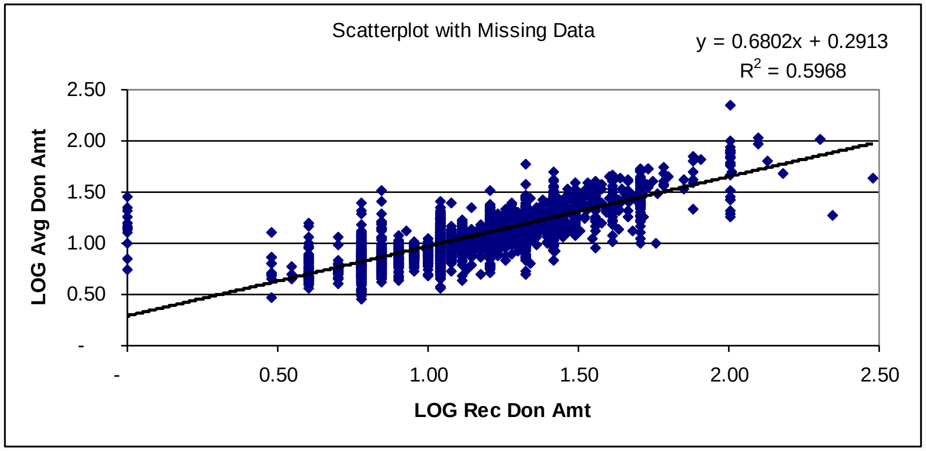
- Missing data can appear as
  - blank, NULL, NA, or a code such as 0, 99, 999, or -1.
- Fixing Missing Data:
  - Delete the record (row), or delete the field (column)
  - Replace missing value with mean, median, or distribution
  - Replace with the missing value with an estimate
    - Select value from another field having high correlation with variable containing missing values
    - Build a model with variable containing missing values as output, and other variables without missing values as an input
- Other considerations
  - Create new binary variable (1/0) indicating missing values
  - Know what algorithms and software do by default with missing values
    - Some do listwise deletion, some recode with "0", some recode with midpoints or means

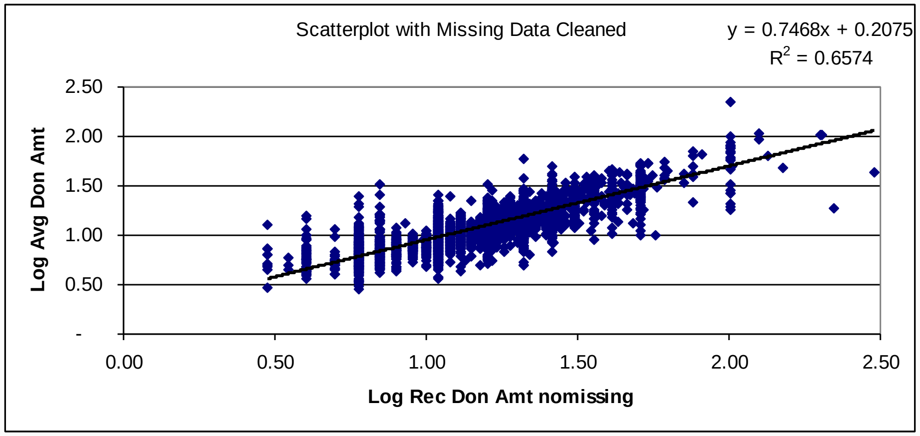
Missing Data:
Imputation with Mean vs. Distribution

payments	original data (no missing)	Cumulative with 10% missing	Cumulative with 10% missing, recoded	Cumulative with 10% missing, recoded	Cumulative with 30% missing	Cumulative with 30% missing, recoded	Cumulative with 30% missing, recoded
500	500	500	500	500	500	500	500
2,019	2,519	2,519	2,519	2,519	2,519	2,519	2,519
7,528	10,047	10,047	10,047	10,047	10,047	10,047	10,047
7,954	18,001	18,001	18,001	18,001	18,001	18,001	18,001
8,438	26,439	26,439	26,439	26,439		**80,737**	**5,200**
8,917	35,356	35,356	35,356	35,356	35,356	35,356	35,356
9,471	44,827		**88,334**	**170,000**		**80,737**	**170,000**
9,912	54,739	54,739	54,739	54,739	54,739	54,739	54,739
10,373	65,112	65,112	65,112	65,112	65,112	65,112	65,112
10,930	76,042	76,042	76,042	76,042	76,042	76,042	76,042
11,392	87,434	87,434	87,434	87,434	87,434	87,434	87,434
11,855	99,289	99,289	99,289	99,289		**80,737**	**160,000**
12,357	111,646	111,646	111,646	111,646	111,646	111,646	111,646
12,862	124,508	124,508	124,508	124,508	124,508	124,508	124,508
13,340	137,848	137,848	137,848	137,848		**80,737**	**22,222**
13,856	151,704	151,704	151,704	151,704	151,704	151,704	151,704
14,252	165,956		**88,334**	**37,000**		**80,737**	**37,000**
14,813	180,769	180,769	180,769	180,769	180,769	180,769	180,769
15,351	196,120	196,120	196,120	196,120		**80,737**	**125,000**
15,817	211,937	211,937	211,937	211,937	211,937	211,937	211,937
Mean	**90,040**	**88,334**	**88,334**	**89,851**	**80,737**	**80,737**	**82,487**
Std. Dev.	**67,415**	**67,949**	**64,274**	**67,959**	**67,745**	**56,037**	**67,611**
Median	**81,738**	**81,738**	**87,884**	**81,738**	**70,577**	**80,737**	**70,577**
Min	**500**	**500**	**500**	**500**	**500**	**500**	**500**
Max	**211,937**	**211,937**	**211,937**	**211,937**	**211,937**	**211,937**	**211,937**
			MAINTAIN MEAN	*MAINTAIN STDEV about mean*		*MAINTAIN MEAN*	*MAINTAIN STDEV about mean*

Clean Data: Missing Data

- How much can missing data effect models?
- Example at upper right has 5300+ records, 17 missing values encoded as "0"
- After fixing model with mean imputation, R^2 rises from 0.597 to 0.657
- Why? Missing was recoded with "0" in this example, which was a particularly bad imputation for this data





Transforms: Changing Distribution of Data

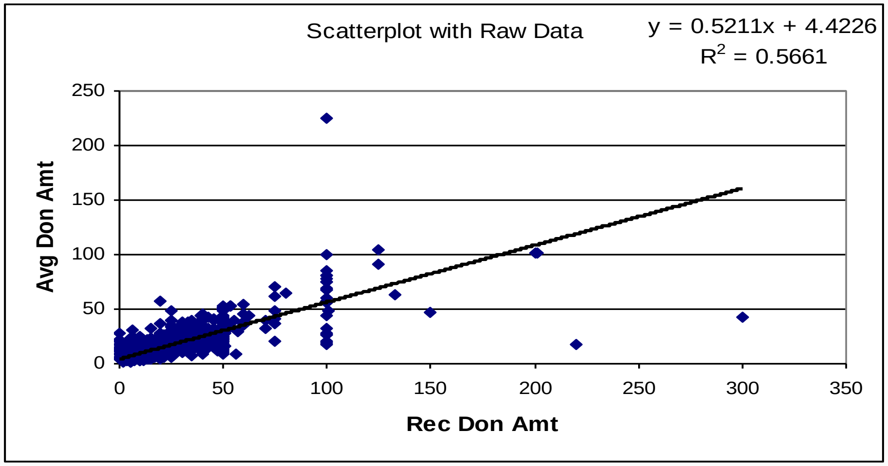
- Positive Skew
  - Tail of distribution to right
  - Correction: log transform
  - Example: MAX_DON_AMT
- Negative Skew
  - Tail of distribution to left
  - Correction: Power >= 2, Exp
  - Example: HOMEVAL50

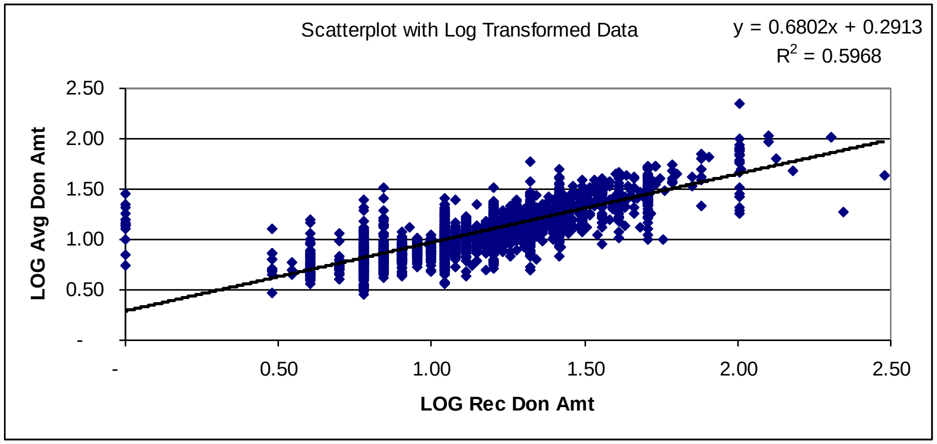
Cells contain: Mean (Standard Deviation)

Field	Selected Transform	Current Distribution	Inverse	LogN	Log10	Exponential	Square Root
AGE	Current Distribution	61.506 (14.419)	0.017 (0.005)	4.089 (0.253)	1.776 (0.110)	325631532221583O...	7.786 (0.944)
HOMEVAL50	Current Distribution	75.193 (28.453)	0.023 (0.058)	4.181 (0.657)	1.816 (0.285)	3020321098838011...	8.311 (2.236)
NUMGIFT_LIFE	Current Distribution	9.402 (8.513)	0.263 (0.289)	1.829 (0.975)	0.794 (0.423)	834022260446348O...	2.789 (1.274)
REC_DON_A...	Current Distribution	17.887 (12.461)	0.077 (0.054)	2.733 (0.556)	1.187 (0.241)	3605097244323034...	4.062 (1.180)
AVG_DON	Current Distribution	13.744 (8.630)	0.095 (0.053)	2.484 (0.515)	1.079 (0.224)	9656746729556163...	3.580 (0.964)

Why Skew Matters (In Regression Modeling)

- Obscures information in plot
  - Spaced in scatterplot taken up by empty space in upper (or lower) end of skewed values
- Regression models fit worse with skewed data
  - In example at right, by simply applying the log transform, performance is improved from R^2=0.566 to 0.597

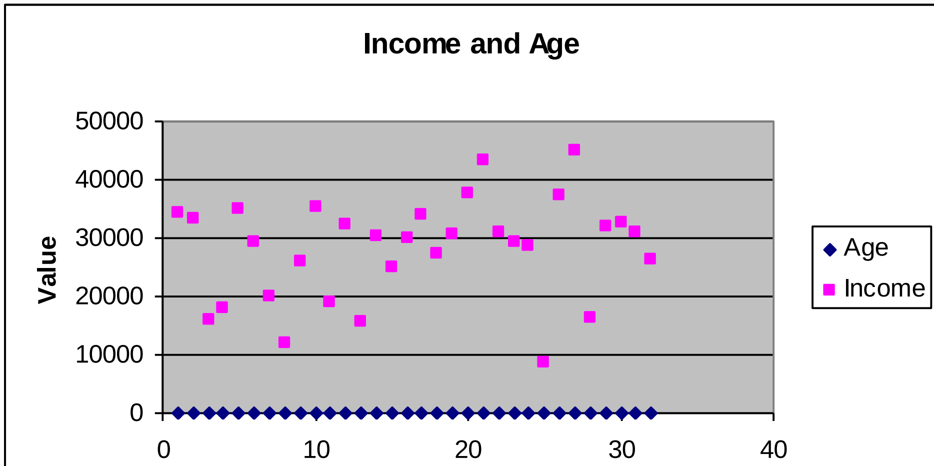


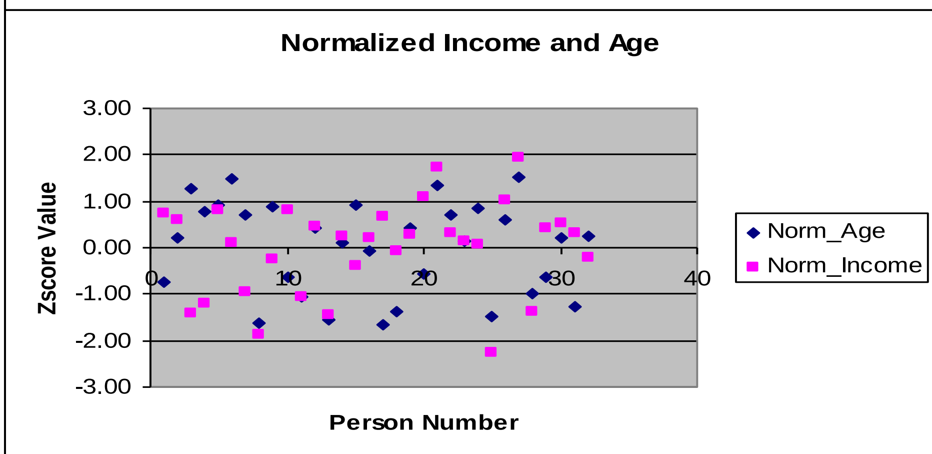


To KNIME

KNIME outlier

Transforms: Scaling Data





- Before normalization, income scale "dwarfs" age
- z-score
  - $x^* = (x - \text{mean}) / \text{std}$
  - Income and age on same scale
- Scale to range [0,1]
  - $x^* = (x - x_{min}) / (x_{max} - x_{min})$
- Both allow one to see both variables on same scale
- Can apply this to subsamples of data (regional data, for example)

Grouping and Exploding Categorical Data

Group: State to Region

State	Group
AK	Northwest
AL	Southeast
AR	Southeast
CA	Southwest
CO	Mountain
CT	Northeast
DC	Mid-Atlantic
DE	Mid-Atlantic
FL	Southeast
GA	Southeast
HI	Southwest
IA	Midwest
ID	Northwest
IL	Midwest
IN	Midwest

Explode: Region "dummy" variables

Region	Mid-Atlantic	Midwest	Mountain	Northeast	Northwest	Southeast
Northwest	0	0	0	0	1	0
Southeast	0	0	0	0	0	1
Southeast	0	0	0	0	0	1
Southwest	0	0	0	0	0	0
Mountain	0	0	1	0	0	0
Northeast	0	0	0	1	0	0
Mid-Atlantic	1	0	0	0	0	0
Mid-Atlantic	1	0	0	0	0	0
Southeast	0	0	0	0	0	1
Southeast	0	0	0	0	0	1
Southwest	0	0	0	0	0	0
Midwest	0	1	0	0	0	0
Northwest	0	0	0	0	1	0
Midwest	0	1	0	0	0	0
Midwest	0	1	0	0	0	0

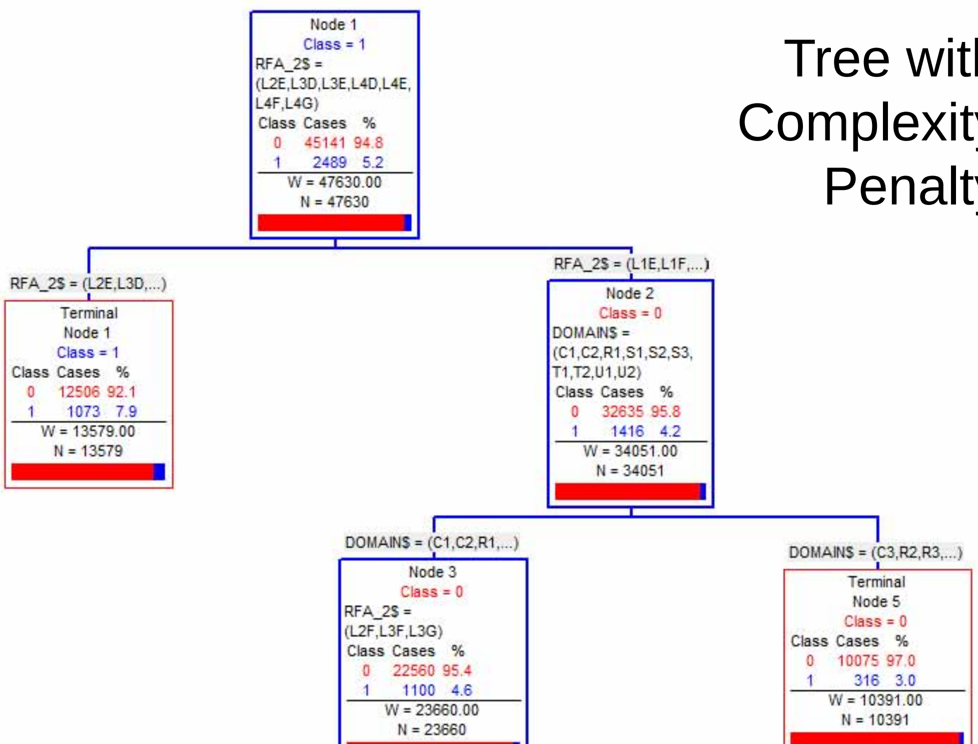
- **Categorical data having values with small populations (10s of cases) is *very* problematic in modeling. They should be binned up (grouped) as much as is possible!**

Effect of High Cardinality

- Cardinality: number of levels in a variable
  - We care about cardinality in categorical variables
  - # levels -> frequency counts
- Why do we care?
  - Decision trees are biased toward accepting splits with variables having high cardinality
  - Numeric algorithm implementations that automatically create dummy variables for categoricals may create *lots* of new 1/0 dummies
    - Higher # inputs in models
    - Lots of low information content variables

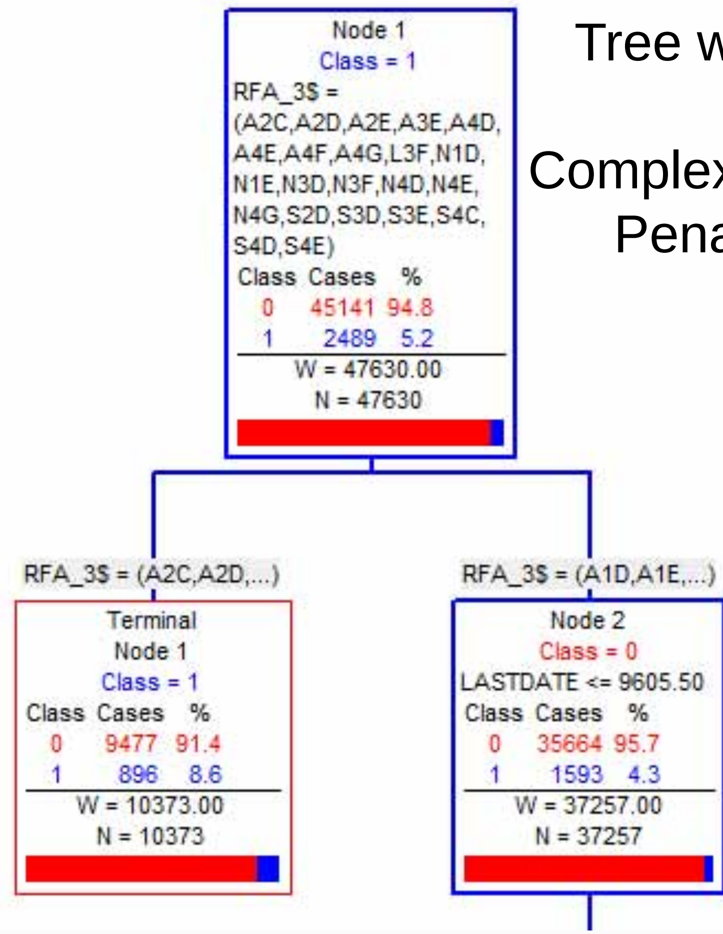
Effect of High Cardinality

Tree with Complexity Penalty

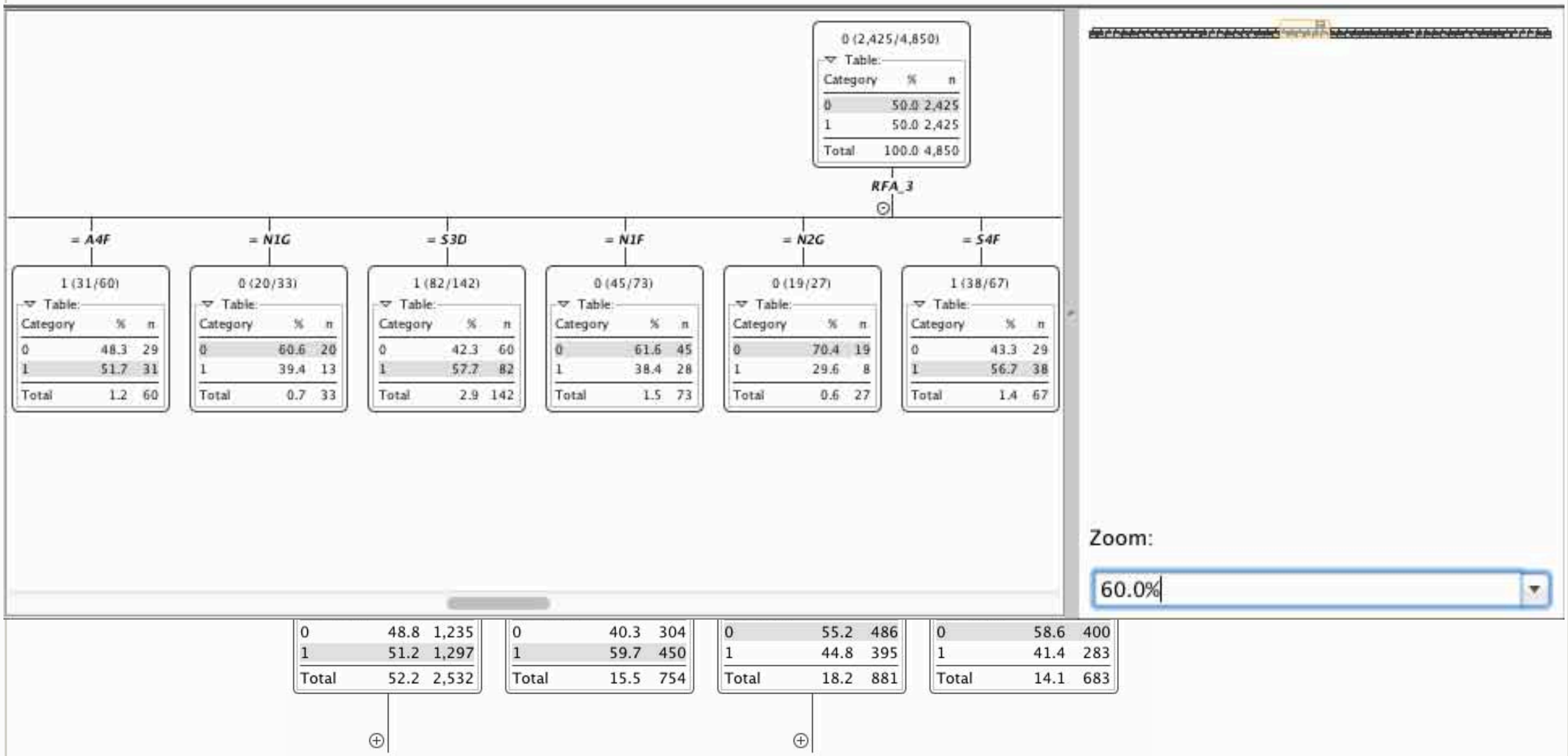


Variable	# populated	# missing	% missing	# levels
RFA_2$	95412	0	0	14
RFA_3$	93462	1950	2.04	70

Tree with no Complexity Penalty



KNIME Decision Trees and High Cardinality



To KNIME

KNIME cardinality

Strategy 5: Deploy Models Wisely

- Get the data right
- Understand how algorithms can be fooled with "correct" data
  - Outliers
  - Missing Values
  - Skew
  - High Cardinality

What is Deployment?

- Using the model for its intended purpose
  - Reporting
  - Transactional scoring
  - Batch scoring
- Most models relate to decisions to be made within the organization

Different Approaches to Deployment

Data Prep	Model	Type of Application
In Database	In Database	real-time scoring
	In PA Software	weekly/monthly scoring
	Standalone	real-time scoring
	In Cloud	large, real-time scoring
in PA Software	In Database	large, real-time scoring
	In PA Software	ad hoc scoring
	Standalone	complex prep; occasional scoring
	In Cloud	big data; complex prep; occasional scoring
In Cloud	In Database	big data; complex prep; occasional scoring
	In PA Software	unlikely
	Standalone	big data; complex prep; occasional scoring
	In Cloud	large, real-time scoring; computationally intensive prep

Form of Models for Deployment: In-PA Software Deployment

- Run models through original software in ad hoc or automated process
- Benefits:
  - Data prep done in software still there
    - But still may have to trim down processing for efficiency
  - no further work to be done to deploy
- Drawbacks
  - Usually slower
    - have to pull data out and push it back to database
    - Software not usually optimized for speed; optimized for usability
  - Requires a software expert to maintain and troubleshoot
    - Analyst usually involved
    - Errors not always handled gracefully

Form of Models for Deployment: External Call to PA Software

- Run models through original software in ad hoc or automated process, but as a call from the OS
- Benefits:
  - Data prep done in software still there
    - But still may have to trim down processing for efficiency
  - no further work to be done to deploy
- Drawbacks
  - Usually slower
    - have to pull data out and push it back to database
    - Software not usually optimized for speed; optimized for usability
  - Requires a software expert to maintain and troubleshoot
    - Analyst usually involved
    - Errors not always handled gracefully

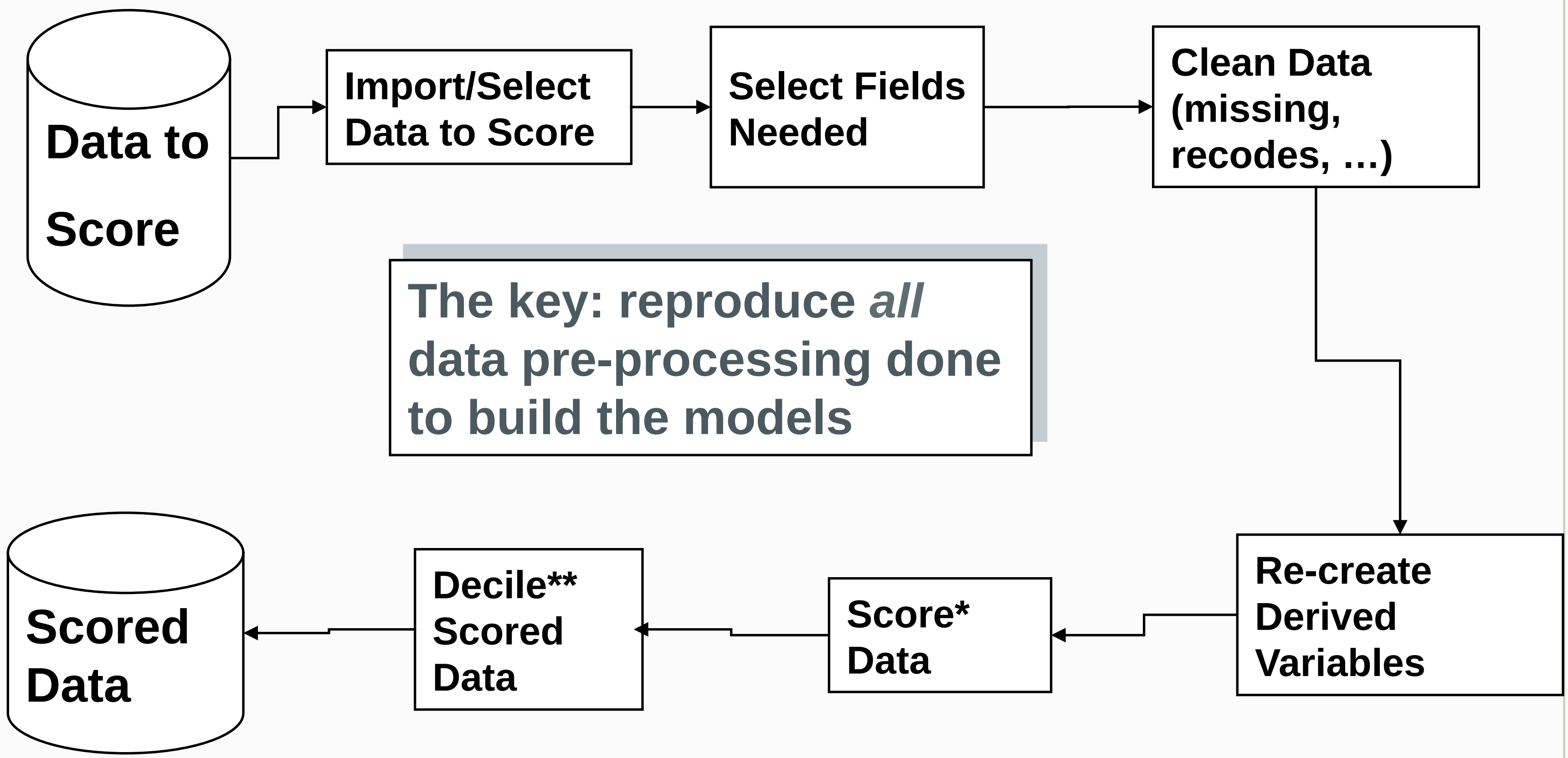
Form of Models for Deployment: Translation to Another Language

- Translate models into SQL, C (++, #, etc.), Java, PMML
  - If in C/Java, can create standalone application just for the model scoring
- Benefits
  - Get models out of software environment where they can be run and maintained by others
  - Often run more efficiently in database or other environment
  - Many tools provide export capabilities into other languages
- Drawbacks
  - Translation of dataprep not usually included in tool export, requires significant time and QC/QA to ensure consistency with the tool
  - Bug fixes take longer

Form of Models for Deployment: PMML

- Translate models into PMML
  - Different than SQL, C, Java, etc.
- Benefits
  - PMML supports (natively) entire predictive modeling process
  - Language is simple
  - Database support
  - Online support for scalable scoring (Zementis)
- Drawbacks
  - Translation of dataprep not usually included in predictive modeling software tools, requires coding
  - Models are verbose
  - Open source scoring options are limited

Typical Model Deployment Processing Flow



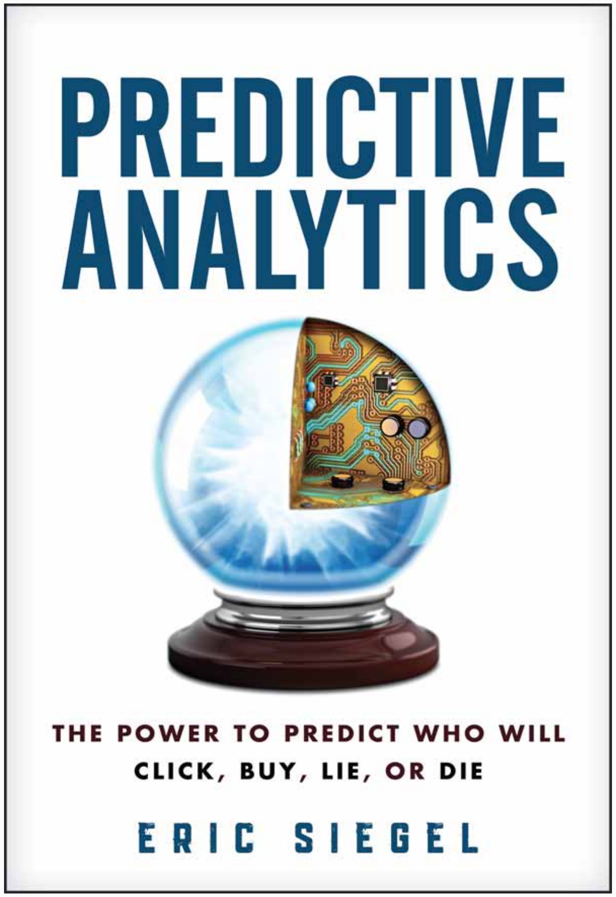
Resources

Dean Abbott
Abbott Analytics, Inc.
Predictive Analytics World, Berlin (#pawcon)
November 6, 2013

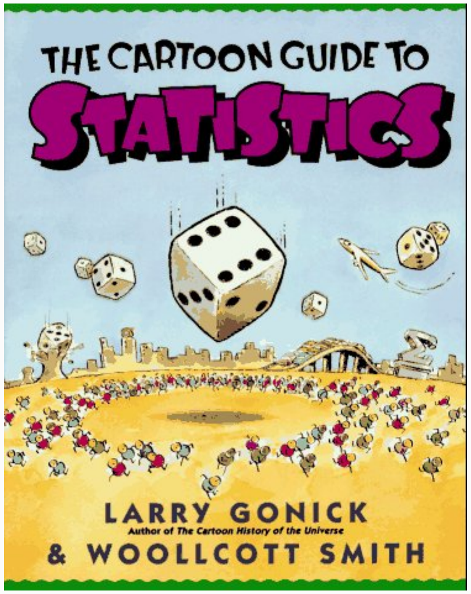
Email: dean@abbottanalytics.com
Blog: http://abbottanalytics.blogspot.com
Twitter: @deanabb

Predictive Analytics Overview

- From Amazon.com
  - Hardcover: 320 pages
  - Publisher: Wiley; 1 edition (February 18, 2013)
  - Language: English
  - ISBN-10: 1118356853
  - ISBN-13: 978-1118356852
- Great introduction to Predictive Analytics

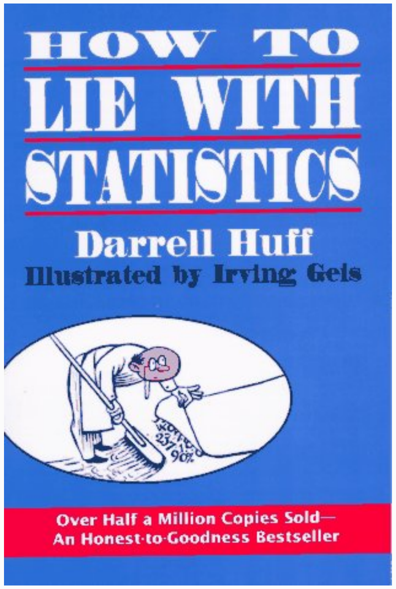


Accessible Books on Statistics



•**The Cartoon Guide to Statistics**
by Larry Gonick, Woollcott Smith
(Contributor), Woollcott Smith

Paperback - 240 pages (February 25, 1994)
HarperCollins (paper); ISBN: 0062731025

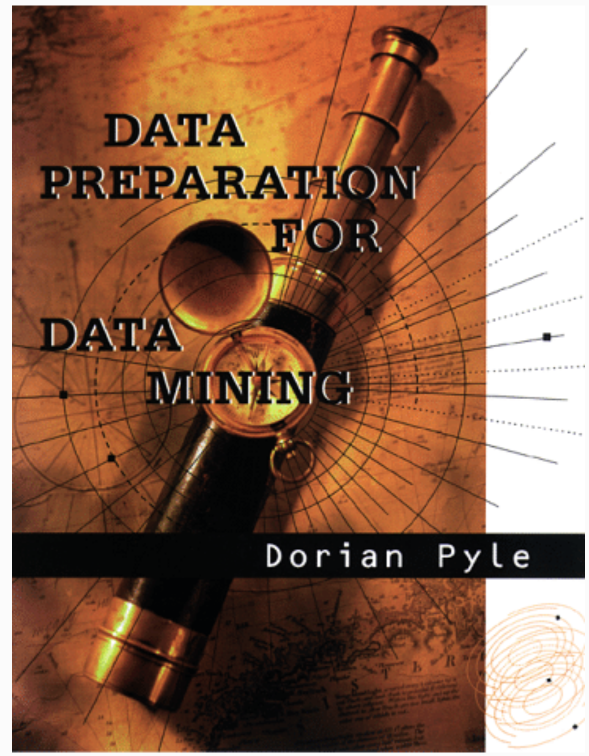


How to Lie With Statistics
by Darrell Huff, Irving Geis (Illustrator)

Paperback Reissue edition (November 1993)
W.W. Norton & Company; ISBN: 0393310728

Data Preparation

- From Amazon.com
  - Data Preparation for Data Mining
  - by Dorian Pyle
  - Paperback - 540 pages Bk&Cd Rom edition (March 15, 1999)
  - Morgan Kaufmann Publishers;
  - ISBN: 1558605290 ;



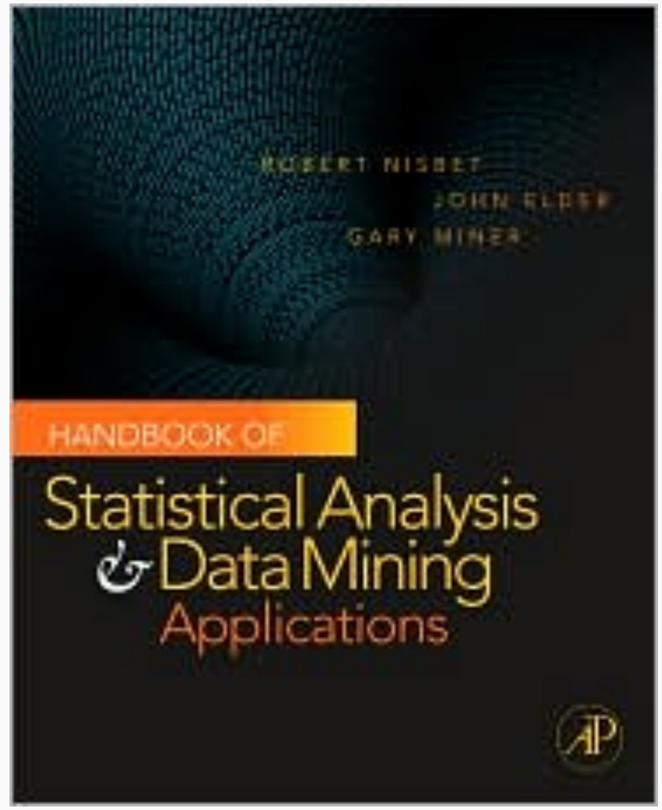
- Excellent resource for the part of data mining that takes the most time. Best book on the market for data preparation.

Data Mining Methods

From Amazon.com

- Handbook of Statistical Analysis and Data Mining Applications
  by Robert Nisbet, John Elder, Gary Miner

- Hardcover: 900 pages
- Publisher: Academic Press (April 23, 2009)
- Language: English
- ISBN-10: 0123747651
- ISBN-13: 978-0123747655



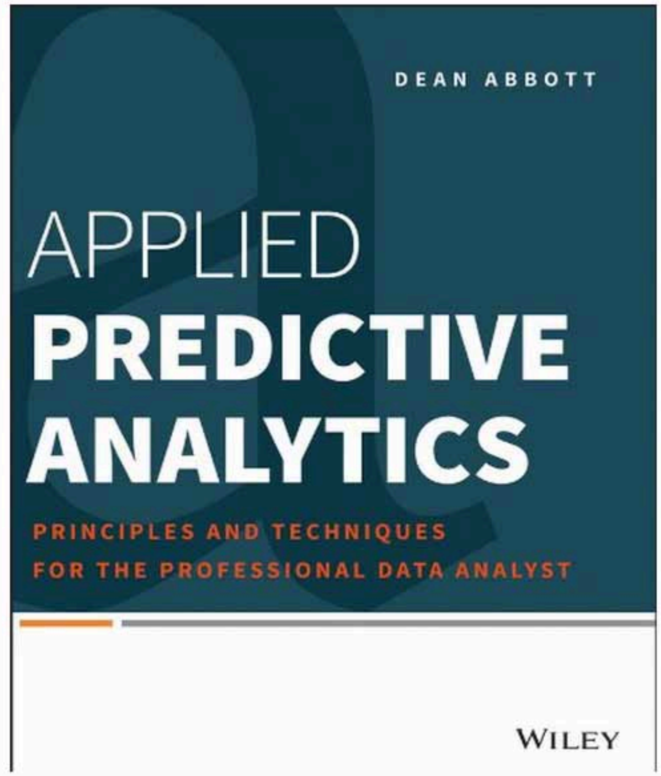
New data mining book written for practitioners, with case studies and specifics of how problems were worked in Enterprise Miner, Clementine, STATISTICA, or another tool

66

Applied Predictive Analytics

Learn the art and science of predictive analytics — techniques that get results

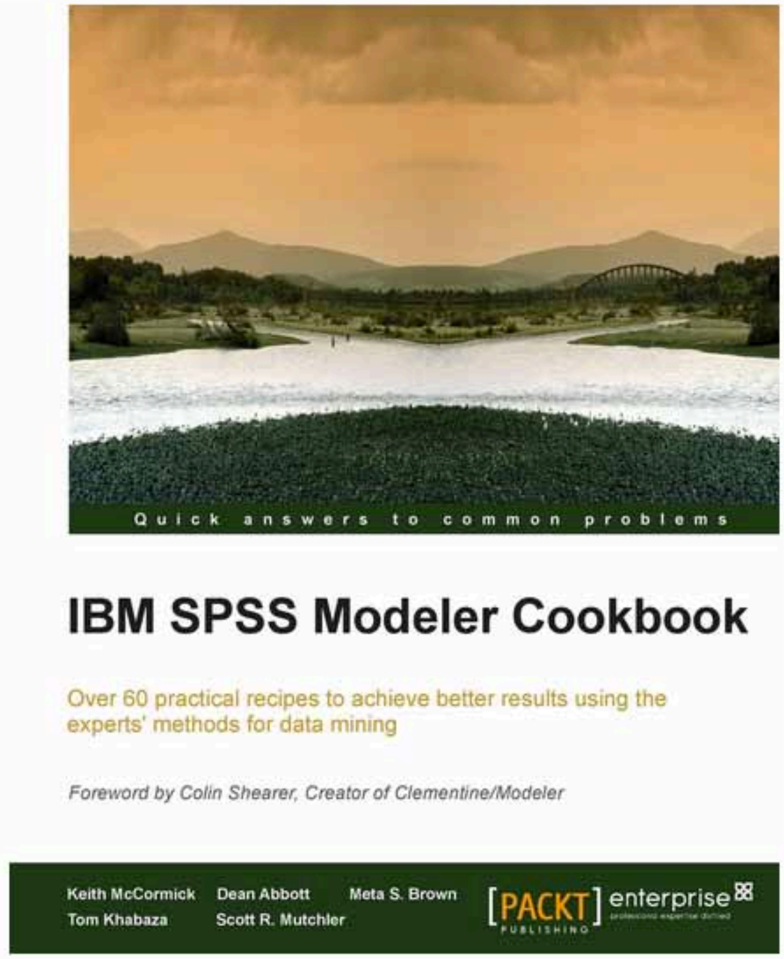
Predictive analytics is what translates big data into meaningful, usable business information. Written by a leading expert in the field, this guide examines the science of the underlying algorithms as well as the principles and best practices that govern the art of predictive analytics. It clearly explains the theory behind predictive analytics, teaches the methods, principles, and techniques for conducting predictive analytics projects, and offers tips and tricks that are essential for successful predictive modeling. Hands-on examples and case studies are included.



Publication Date: March 31, 2014 | ISBN-10: 1118727967 | ISBN-13: 978-1118727966 | Edition: 1

IBM Modeler Recipes

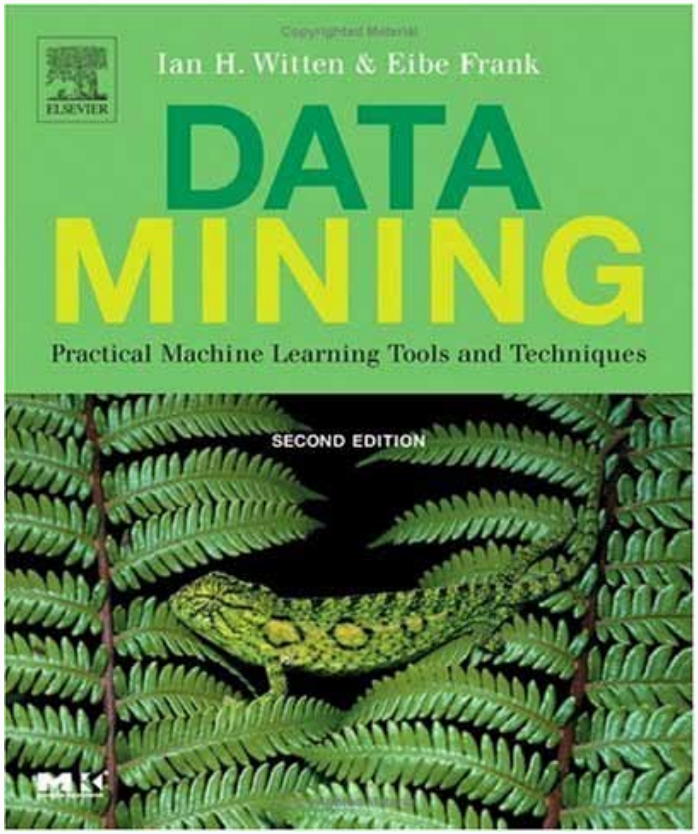
Go beyond mere insight and build models than you can deploy in the day to day running of your business
Save time and effort while getting more value from your data than ever before
Loaded with detailed step-by-step examples that show you exactly how it's done by the best in the business



Book Details
Language : English
Paperback : 386 pages [235mm x 191mm]
Release Date : November 2013
ISBN : 1849685460
ISBN 13 : 9781849685467
Author(s) : Keith McCormick, Dean Abbott, Meta S. Brown, Tom Khabaza, Scott Mutchler
Topics and Technologies : All Books, Cookbooks, Enterprise

Data Mining Algorithms

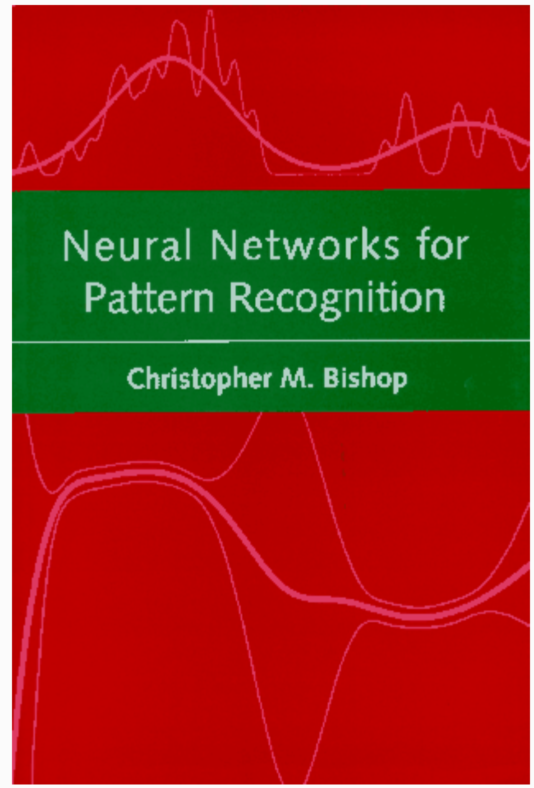
- From Amazon.com
  - Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations
  - By Eibe Frank, Ian H. Witten
  - Paperback - 416 pages (October 13, 1999)
  - Morgan Kaufmann Publishers;
  - ISBN: 1558605525;
- Best book I've found in between highly technical and introductory books. Good coverage of topics, especially trees and rules, but no neural networks.



Data Mining Algorithms

From Amazon.com

- Neural Networks for Pattern Recognition by Christopher M. Bishop
- Paperback (November 1995)
- Oxford Univ Press;
- ISBN: 0198538642



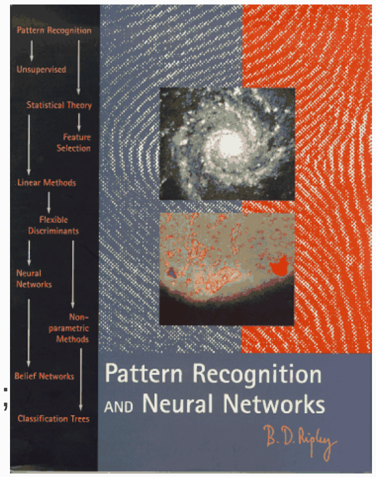
Excellent book for neural network algorithms, including some lesser known varieties.

Described as "Best of the best" by Warren Sarle (Neural Network FAQ)

Data Mining Algorithms

From Amazon.com

- Pattern Recognition and Neural Networks
- by Brian D. Ripley,
- N. L. Hjort (Contributor)
- Hardcover (October 1995)
- Cambridge Univ Pr (Short);
- ISBN: 0521460867



Ripley is a statistician who has embraced data mining. This book is not just about neural networks, but covers all the major data mining algorithms in a very technical and complete manner.

Sarle calls this the best advanced book on Neural Networks

From Amazon.com

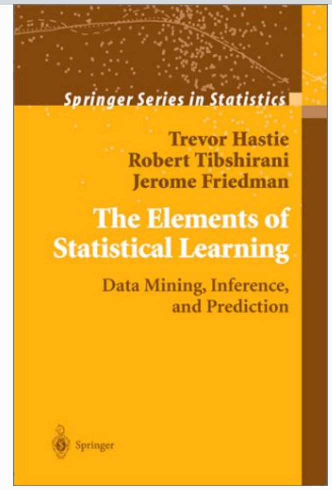
The Elements of Statistical Learning

by Trevor Hastie, Rob Tibsharani Jerome Friedman

Hardcover (2001)

Springer;

ISBN: 0-387-95284-5

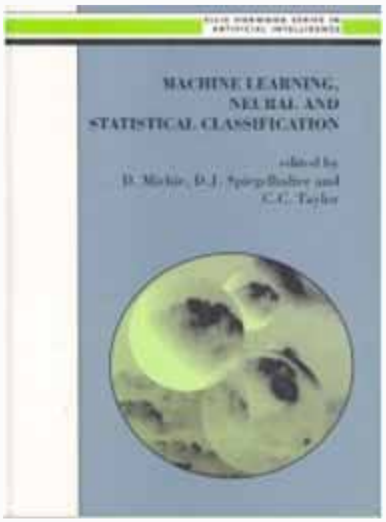


By 3 giants of the data mining community, I have read most of the book and can't think of a significant conclusion I disagree with them on. Very technical, but very complete. Topics covered in this book not usually covered in others such as kernel methods, support vector machines, principal curves, and many more. Has become my favorite technical DM book.

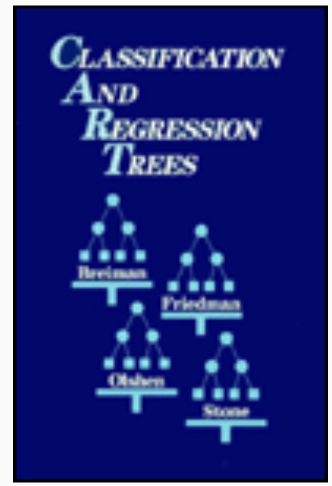
Book has 200 color figures/charts—first data mining book I've seen that makes use of color, and this book does it right

http://www-stat.stanford.edu/~tibs/ElemStatLearn/download.html

Accessible Technical Description of Algorithms

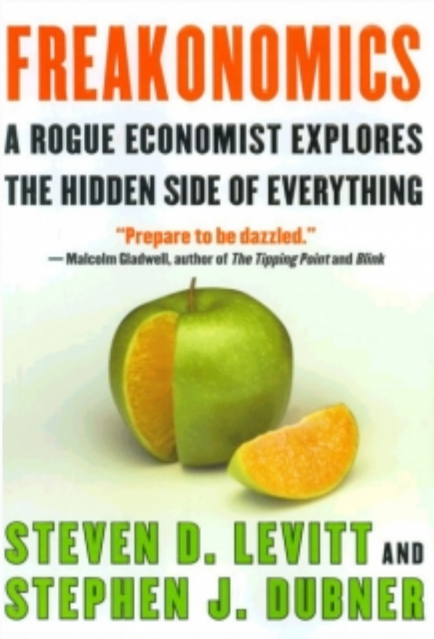


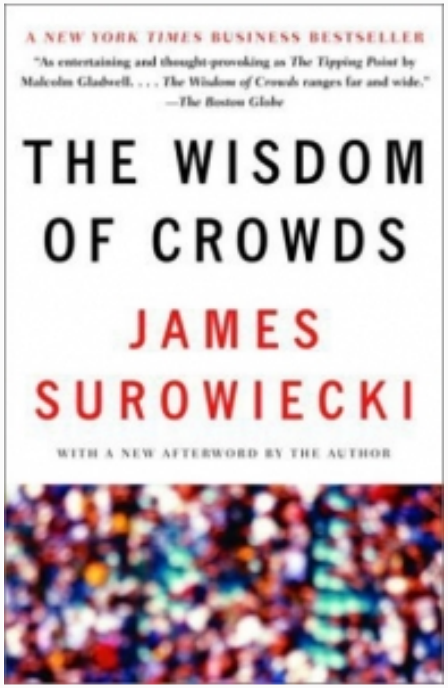
- Machine Learning, Neural and Statistical Classification
- D. Michie, D.J. Spiegelhalter, C.C. Taylor (eds)
- Available free online (PDF)
- http://www.amsta.leeds.ac.uk/~charles/statlog/

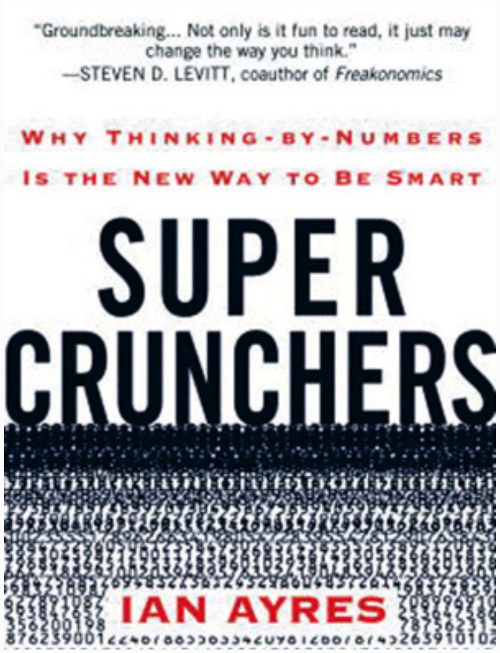


- From Amazon.com
- Classification and Regression Trees by Leo Breiman
- Paperback (December 1983)
- CRC Press;
- ISBN: 0412048418
- The definitive textbook on decision trees from the inventors of the CART algorithm.

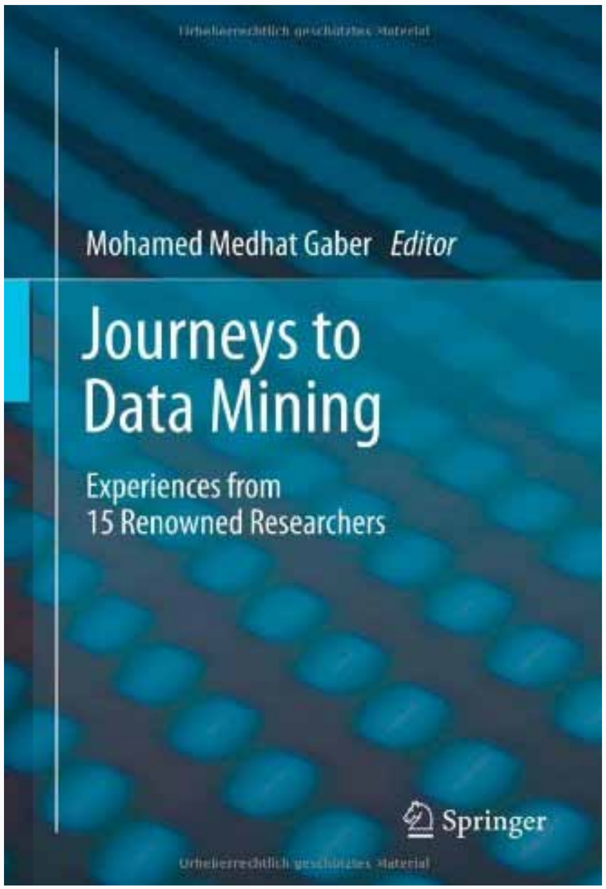
Popular Data Mining Books







Biographies



With contributions from:

Dean Abbott, @deanabb
Charu Aggarwal
Michael Berthold
Chris Clifton
John Elder IV
David J. Hand
Cheryl G. Howard
J. Dustin Hux
Hillol Kargupta
Colleen McLaughlin McCue
G. J. McLachlan
Gregory Piatetsky-Shapiro, @kdnuggets
Shusaku Tsumoto
Graham J. Williams
Mohammed J. Zaki

Publication Date: July 21, 2012
ISBN-10: 3642280463
ISBN-13: 978-3642280467

Descriptions of Algorithms

- Neural Network FAQ
- ftp://ftp.sas.com/pub/neural/FAQ.html
- Statistical data mining tutorials by Andrew Moore, Carnegie Mellon
- http://www-2.cs.cmu.edu/~awm/tutorials/
- A list of papers and abstracts from The University of Bonn
  Data Clustering and Visualization is a category of particular interest. Hasn't been updated since 2003, but still a good selection of papers.
- http://www-dbv.informatik.uni-bonn.de
- A Statistical Learning/Pattern Recognition Glossary by Thomas Minka.
  Very comprehensive list of data mining terms and glossary-like descriptions
- http://www.stat.cmu.edu/~minka/statlearn/glossary/