



use R! Group of San Francisco Bay Area

2009 Kickoff Meeting
at
Predictive Analytics World 2009

www.meetup.com/R-Users/



The R and Science of Predictive Analytics: Four Case Studies in R

Panel:

Bo Cowgill, Google

Itamar Rosenn, Facebook

David Smith, Revolution Computing

Jim Porzak, The Generations Network

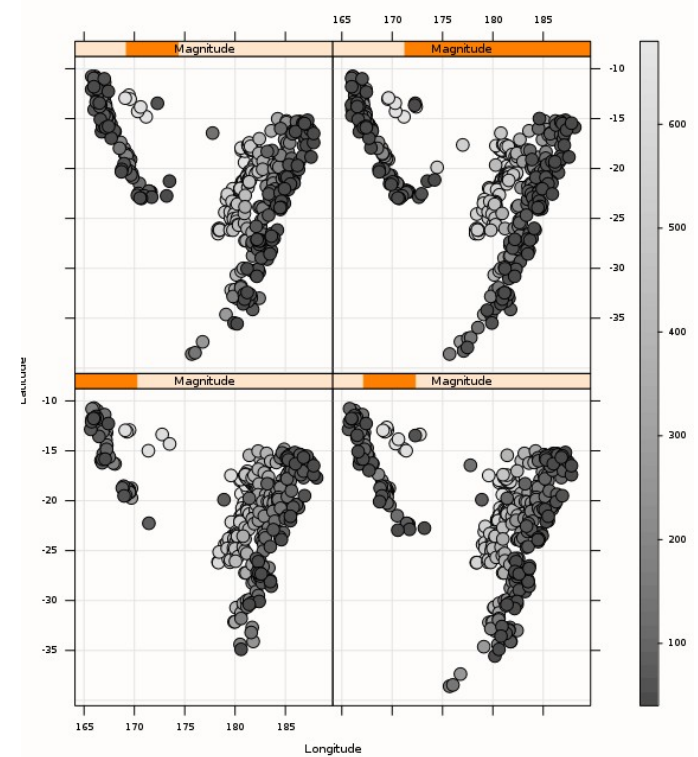
Moderator: Michael Driscoll, Dataspora LLC





What is R?

- A programming language designed for
 - Data manipulation
 - Statistics
 - Data Visualization
- Why sets it apart?
 - Developed by statisticians
 - Free, open source
 - Extensibility via packages





First there was S

- R is the free (GNU), open source, version of S
 - S developed by John Chambers et al while at Bell Labs in 80's
 - For "data analysis and graphics" (with statistics emphasis)
 - Ver. 4 defined by the "Green Book" *Programming with Data*, 1998
- R was initially written in early 1990's
 - by Robert Gentleman and Ross Ihaka
 - Statistics Department of the University of Auckland
 - GNU GPL release in 1995
 - "R" is before "S", as "HAL" is before "IBM"
- Since 1997 a core group of ± 20 developers
 - Since 1997 a core group of ± 20 developers
 - Continually developed with a new 0.1 level release ~ 6 months



A Simple R Example

```
> plot(short.velocity ~ blood.glucose,
      data=thuesen)

> fit <- lm(short.velocity ~ blood.glucose,
            data=thuesen)

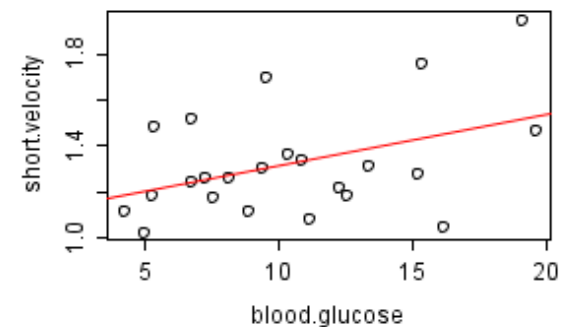
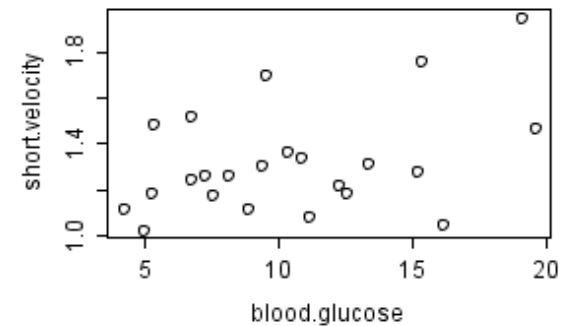
> summary(fit)

Call:
lm(formula = short.velocity ~ blood.glucose, data = thuesen)
Residuals:
    Min       1Q   Median       3Q      Max
-0.40141 -0.14760 -0.02202  0.03001  0.43490

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.09781    0.11748   9.345 6.26e-09 ***
blood.glucose  0.02196    0.01045   2.101  0.0479 *
---
Residual standard error: 0.2167 on 21 degrees of freedom
Multiple R-squared:  0.1737, Adjusted R-squared:  0.1343
F-statistic: 4.414 on 1 and 21 DF,  p-value: 0.0479

> abline(fit, col = "red")

> predict(fit, data.frame(blood.glucose = 15))
1
1.427253
```





Current State of R

As of October, 2004

- V2.0 Released October, 2004
- Windows, Mac, Linux & Unix ports
- Over 400 submitted packages from "abind" to "zoo"
- 12th newsletter (Volume 4/2) published September 2004
- The first useR! - R User Conference held in Vienna May 2004
- ~400 R-help messages per week
- ~ Dozen texts specifically on R or with R examples and code
- R language generally accepted to be more powerful than S-Plus
- Some interesting GUI work in progress

As of February, 2009

- V2.8.1 Released December, 2008
- Vista, Ubuntu, 64bit versions
- 1697 packages; "ADaCHG" to "zyp" (+37 Omega, +296 Bioconductor)
- 23rd Newsletter (Vol. 8/2), October 2008
- 5th useR! this July in Rennes, France
- ~ 700 R-help messages per week
- 74 texts now listed on r-project, including: *Software for Data Analysis, Programming with R* by John Chambers
- R ~ universally taught & used academically for development
- JGR, Rattle, RCmdr, ...
- Interesting large application work in progress including R in the clouds



Finding Prediction Methods in R

- CRAN Task Views for quick guide to packages:
 - Machine Learning & Statistical Learning
<http://cran.cnr.berkeley.edu/web/views/MachineLearning.html>
 - Multivariate Statistics
<http://cran.cnr.berkeley.edu/web/views/Multivariate.html>
 - R News for introductory articles.
 - Search PDFs for “predict(“
 - 33 hits in 13 issues
 - Max Kuhn's caret Package:
Building Predictive Models in R Using the caret Package
www.jstatsoft.org/v28/i05
-



Models supported by caret (1 of 4)

Model	method	Value	Package	Tuning Parameters
<i>“Dual-Use Models”</i>				
Generalized linear model	glm		stats	None
Recursive Partitioning	rpart		rpart	maxdepth
	ctree		party	mincriterion
	ctree2		party	maxdepth
Boosted Trees	gbm		gbm	interaction.depth, n.trees, shrinkage
	blackboost		gbm	maxdepth, mstop
	ada		ada	maxdepth, iter, nu
	Other Boosted Models	glmboost	mboost	mstop
Random Forests	gamboost		mboost	mstop
	rf		randomForest	mtry
Bagged Trees	cforest		party	mtry
	treebag		ipred	None
Neural Networks	nnet		nnet	decay, size
Partial Least Squares	pls		pls, caret	ncomp
Sparse Partial Least Squares	spls		spls, caret	K, eta, kappa
Support Vector Machines (RBF kernel)	svmRadial		kernlab	sigma, C
Support Vector Machines (polynomial kernel)	svmPoly		kernlab	scale, degree, C
Gaussian Processes (RBF kernel)	gaussprRadial		kernlab	sigma
Gaussian Processes (polynomial kernel)	gaussprPoly		kernlab	scale, degree



Models supported by caret (2 of 4)

Model	method	Value	Package	Tuning Parameters
<i>Regression Models</i>				
Linear Least Squares	lm		stats	None
Multivariate Adaptive Regression Splines	earth, mars		earth	degree, nprune
Bagged MARS	bagEarth		caret, earth	degree, nprune
M5 Rules	M5Rules		RWeka	pruned
Elastic Net	enet		elasticnet	lambda, fraction
The Lasso	lasso		elasticnet	fraction
Projection Pursuit Regression	ppr		stats	nterms
Penalized Linear Models Regression Splines	penalized		penalized	lambda1, lambda2
Relevance Vector Machines (RBF kernel)	rvmRadial		kernlab	sigma
Relevance Vector Machines (polynomial kernel)	rvmPoly		kernlab	scale, degree
Supervised Principal Components	superpc		superpc	n.components, threshold



Models supported by caret (3 of 4)

Model	method	Value	Package	Tuning Parameters
<i>Classification Models</i>				
Linear Discriminant Analysis	lda		MASS	None
Quadratic Discriminant Analysis	qda		MASS	None
Stabilised Linear Discriminant Analysis	sllda		ipred	None
Shrinkage Linear Discriminant Analysis	sda		sda	diagonal
Sparse Linear Discriminant Analysis	sparseLDA		sparseLDA	NumVars, lambda
Stepwise Diagonal Discriminant Analysis	sddaLDA, sddaQDA		SDDA	None
Regularized Discriminant Analysis	rda		klaR	lambda, gamma
Mixture Discriminant Analysis	mda		mda	subclasses
Penalized Discriminant Analysis	pda pda2		mda mda	lambda df



Models supported by caret (4 of 4)

Model	method	Value	Package	Tuning Parameters
Flexible Discriminant Analysis (MARS basis)	fda		mda, earth	degree, nprune
Bagged FDA	bagFDA		caret, earth	degree, nprune
Logistic/Multinomial Regression	multinom		nnet	decay
LogitBoost	logitboost		caTools	nIter
Logistic Model Trees	LMT		RWeka	iter
C4.5 decision trees	J48		RWeka	C
Least Squares Support Vector Machines (RBF kernel)	lssvmRadial		kernlab	sigma
k Nearest Neighbors	knn3		caret	k
Nearest Shrunken Centroids	pam		pamr	threshold
Naive Bayes	nb		klaR	usekernel
Generalized Partial Least Squares	gpls		gpls	K.prov
Learned Vector Quantization	lvq		class	k

From Table 1 in Max Kuhn's caret package vignette *The caret Package*: :
<http://cran.cnr.berkeley.edu/web/packages/caret/index.html>



Bo Cowgill, Google



Itamar Rosenn, Facebook



David Smith, Revolution Computing

Revolution R Enterprise

An enhanced, high-performance distribution of R, designed for use in commercial environments.

<http://www.revolution-computing.com>



ParalleIR

Easy-to-use parallel computing with R on multicore workstations and clusters



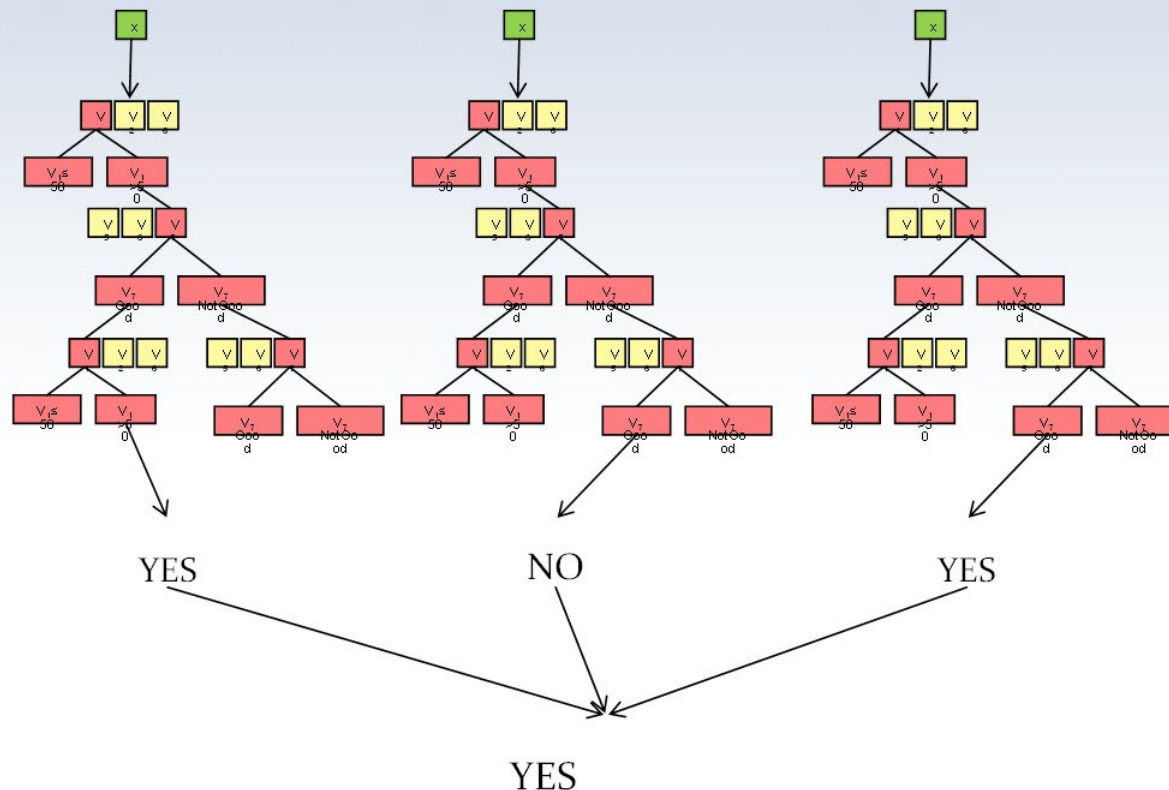
Revolutions blog

News about R, statistics and the world of open-source

<http://blog.revolution-computing.com>

Predicting with Random Forests

1. Build trees with `mtry` random features on bootstrap samples of training data
2. Run the new data down each tree in the forest (independent, parallelizable)
3. Take a majority vote (classification), average (regression), or other single-valued output function of all the tree results



Sequential Implementation

```
library (randomForest)
rf <- randomForest (x, y, ntree=1000)
```

Parallel Implementation with ParallelR

```
library (randomForest)
library (foreach) # from ParallelR 2.0
wc <- workerCount (getSleigh())
n <- ceiling (1000/wc)
rf <- foreach (j=rep(n, wc), COMBINE=combine,
              PACKAGES='randomForest') %dopar%
  randomForest (x, y, ntree=j)
```

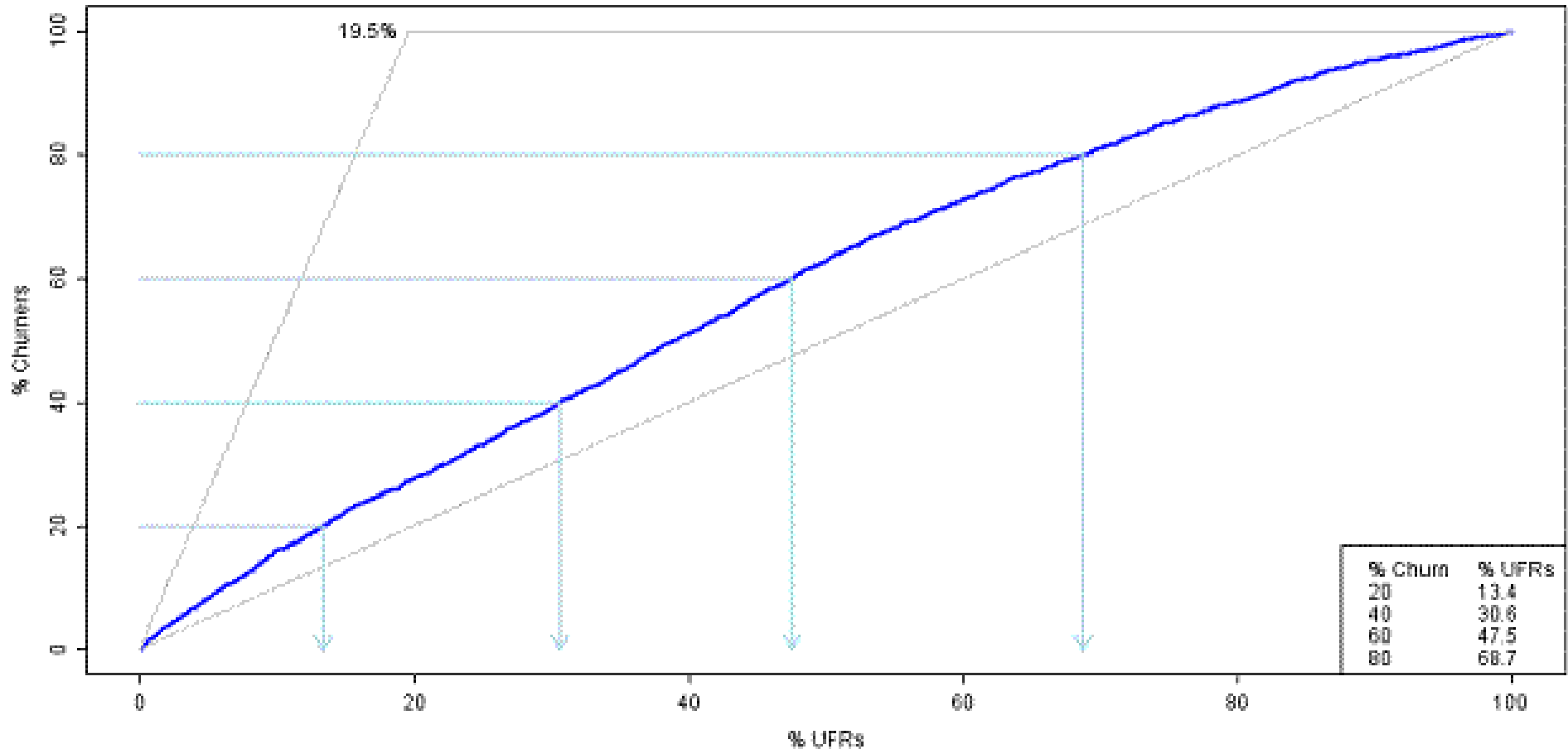


Jim Porzak, The Generations Network



Jim's Marketing Analytics Example

Q1 At-Risk Model Validation



More examples in my talk tomorrow with Alex Kriney from Sun.



Discussion

Q&A



Appendix

Getting Started with R



R Links

- R Homepage: www.r-project.org
 - The official R site
- R Foundation: www.r-project.org/foundation
 - Central reference point for R development
 - Holds copyright / GPL of R software & docs
- Local CRAN: cran.cnr.berkeley.edu
 - Find yours at: cran.r-project.org/mirrors.html
 - Current binaries, docs, FAQs, & more!
- JGR Site: jgr.markushelbig.org/JGR.html



R Basics – Learning More

Wikipedia

[http://en.wikipedia.org/wiki/R_\(programming_language\)](http://en.wikipedia.org/wiki/R_(programming_language))

An Introduction to R

<http://cran.cnr.berkeley.edu/doc/manuals/R-intro.html>

Links to all “official” manuals (html & pdf)

<http://cran.cnr.berkeley.edu/manuals.html>

R Graph Gallery

<http://addictedtor.free.fr/graphiques/>

R Wiki

<http://wiki.r-project.org/rwiki/doku.php>

For SAS & SPSS users (Bob Muenchen's Rosetta Stone)

<http://rforsasandspssusers.com/>