Marrying Prediction and Segmentation to Identify Sales Leads

Jim Porzak, The Generations Network
Alex Kriney, Sun Microsystems

February 19, 2009

Business Challenge

- In 2005 Sun Microsystems began the process of open-sourcing and making its software stack freely available
- Many millions of downloads per month*
- Heterogeneous registration practices
- Multichannel (web, email, phone, in-person) and multitouch marketing strategy
- 15%+ response rates; low contact-to-lead ratio
- **Challenge: Identify the sales leads**

* Solaris, MySQL, GlassFish, NetBeans, OpenOffice, OpenSolaris, xVM Virtualbox, JavaFX, Java, etc.

The Project

- Focus on Solaris 10 (x86 version)
- Data sources: download registration, email subscriptions, other demographics databases, product x purchases
- What are the characteristics of someone with a propensity to purchase?
- How can we become more efficient at identifying potential leads?
- Apply learnings to marketing strategy for all products and track results

Predictive Models Employed

- Building a purchase model using random forests
- Creating prospect persona segmentation using cluster analysis

Random Forest Purchase Model

Random Forests

- Developed by Leo Breiman of Cal Berkeley, one of the four developers of CART, and Adele Cutler, now at Utah State University.
- Accuracy comparable with modern machine learning methods. (SVMs, neural nets, Adaboost)
- Built in cross-validation using “Out of Bag” data. (Prediction error estimate is a by product)
- Large number candidate predictors are automatically selected. (Resistant to over training)
- Continuous and/or categorical predicting & response variables. (Easy to set up.)
- Can be run in unsupervised for cluster discovery. (Useful for market segmentation, etc.)
- Free Prediction and Scoring engines run on PC’s, Unix/Linux & Mac’s. (R version)
- See Appendix for links to more details.

Data for Model

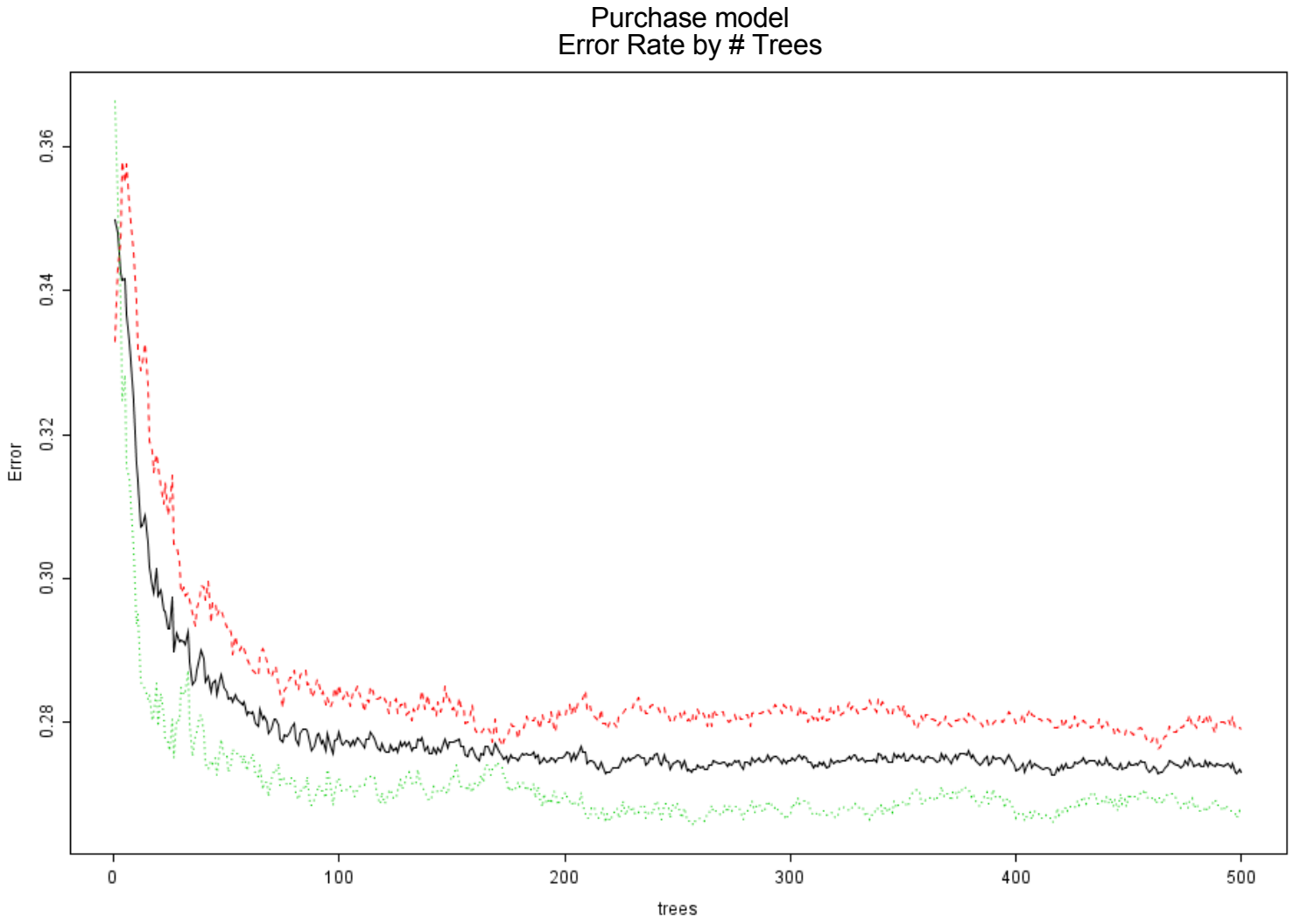
- Data cleanup issues
  - Categorical variables with many categories limited to top 30, with rest grouped into “Other2”
  - Silly answers in number of licenses questions
  - Email domains checked & simplified
- Predict purchase of **product x**; yes or no.
- Types of predictors
  - Demographic: Job & Organization
  - Intended Use by application area
  - Engagement with Sun: Newsletter list combinations
- Solaris product registration *not* used in model – concern that registration was intrinsic in purchase process

randomForest Run

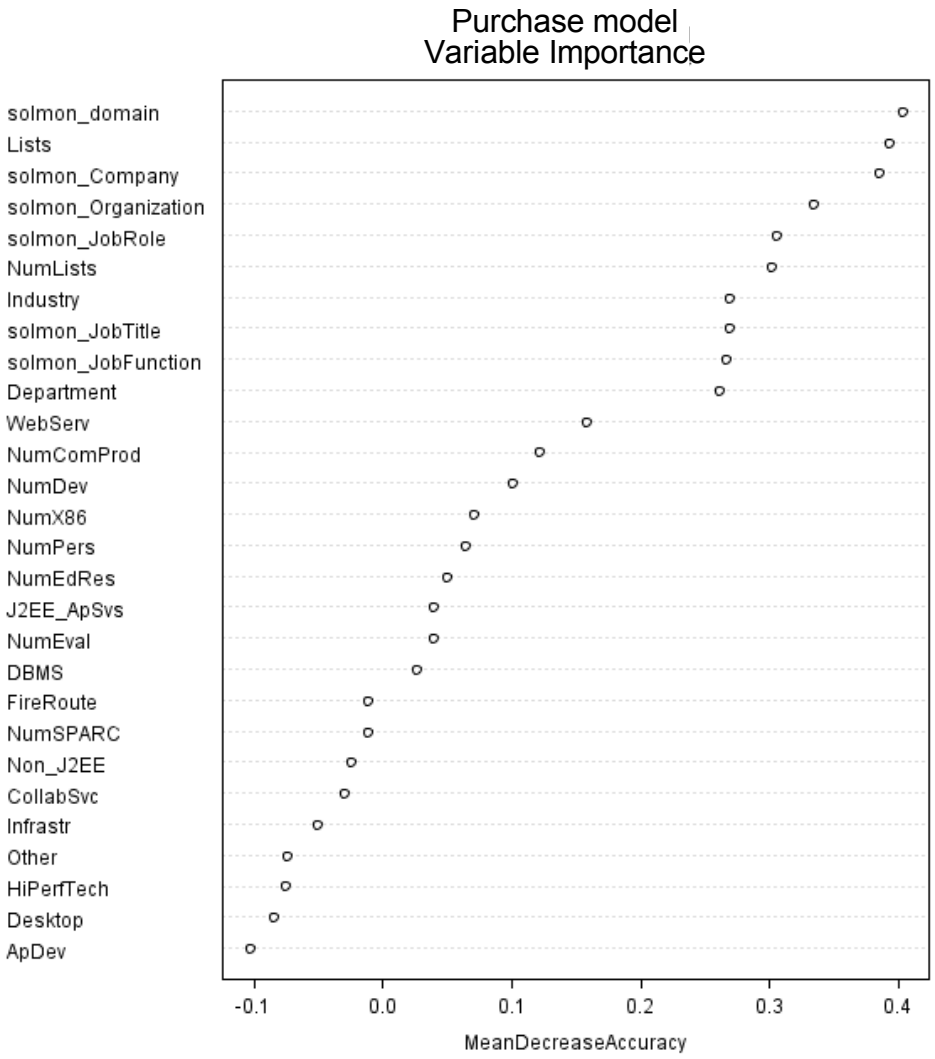
- Training set: random sample of 8000 records
- Tuned & repeated to verify stability.
- Final model:

```
> train2.rf

Call:
 randomForest(x = train2[, -1], y = train2[, 1], classwt = c(5,
      1), importance = TRUE, proximity = FALSE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 5

        OOB estimate of  error rate: 27.3%
Confusion matrix:
          Has Paid Not Paid class.error
Has Paid     2884     1116       0.279
Not Paid     1068     2932       0.267
```

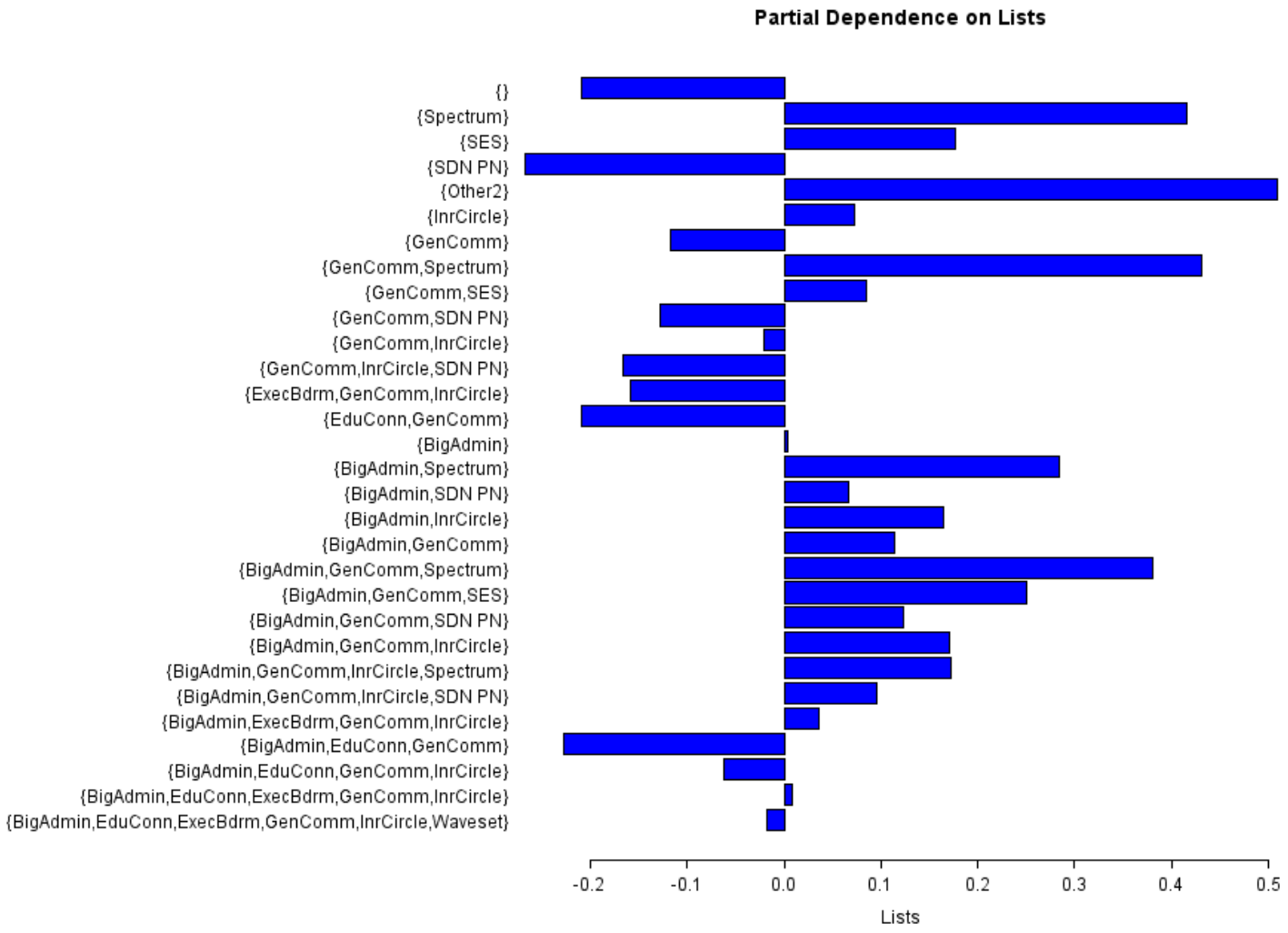
Diagnostics – Error Rate by # Trees



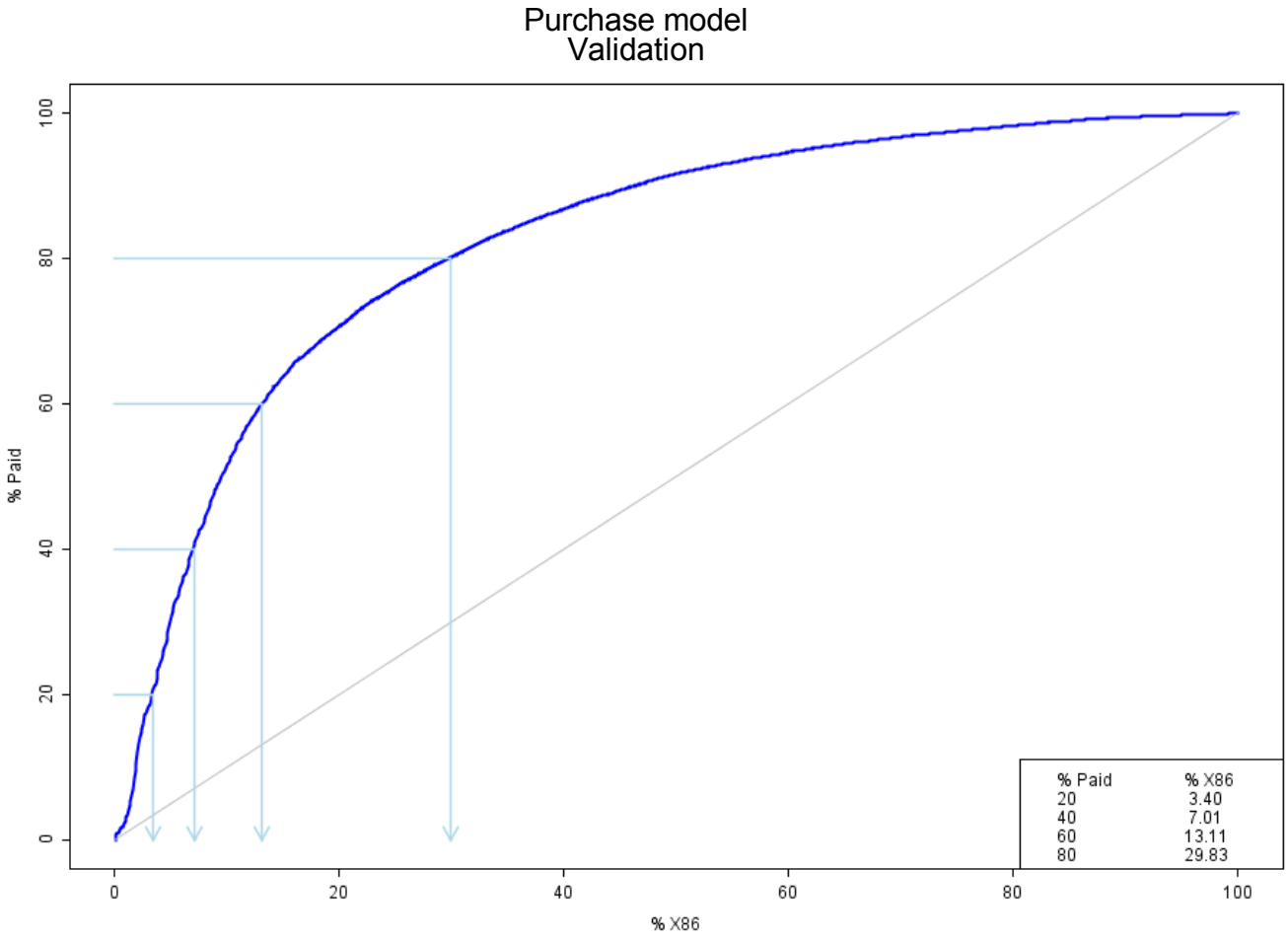
Variable Importance Plot



Partial Dependence Plot Example



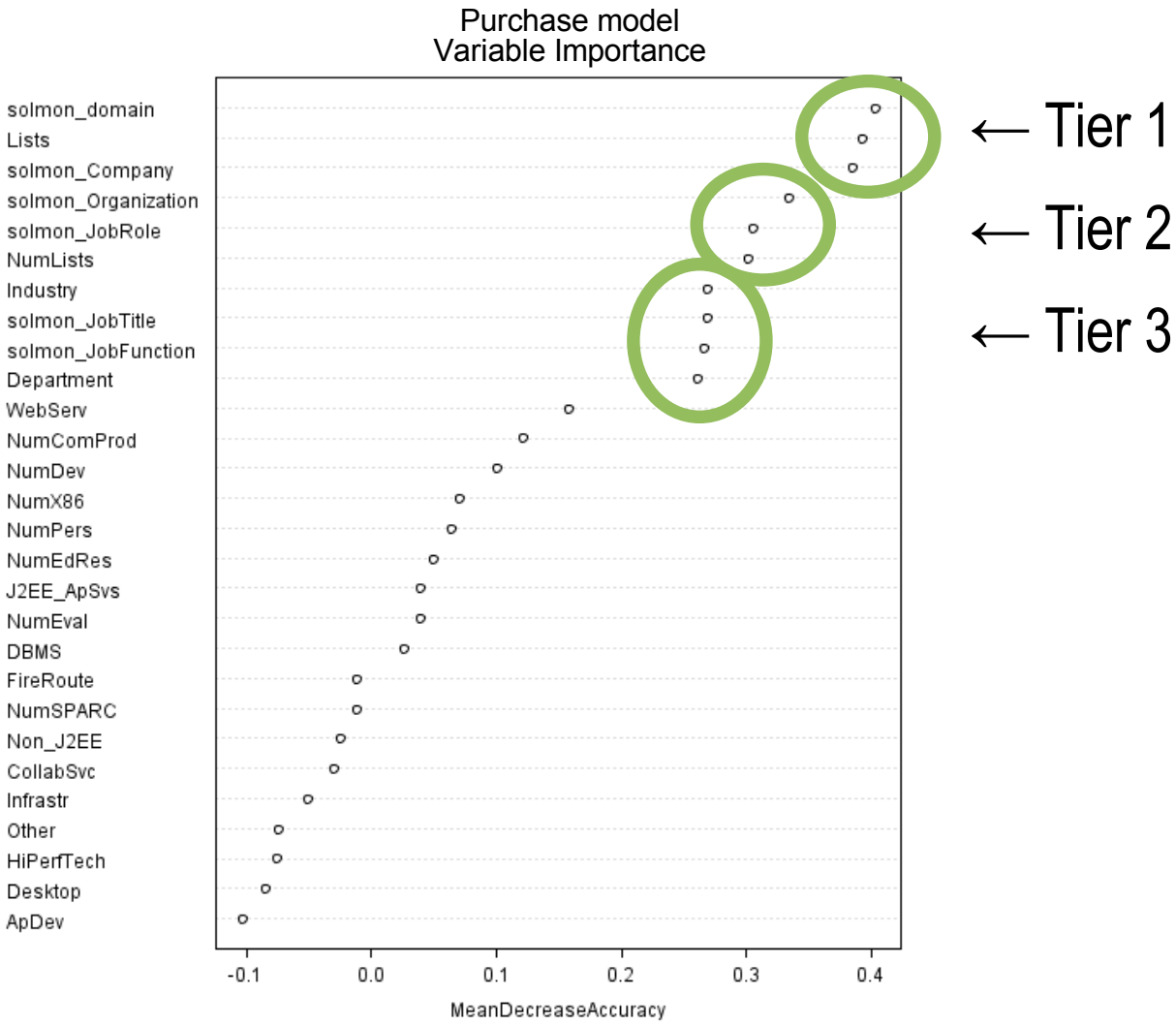
Model worked well!



Trained on 8,000 records, validated on entire database.

% Purchased	% Cumulative depth of file
20%	3.4%
40%	7%
60%	13%
80%	30%

Predictors by Importance



Variables that Predict Purchase

- On the whole those who raise their hands and identify themselves with more detailed company-related information tend to purchase
- Specificity is the key to all the variables
- 3 tiers of variables in order of importance

Tier 1	**Tier 2**	**Tier 3**
- Email domain	- Organization (dept.)	- Industry
- List membership	- Job Role	- Everything else
- Company	- Number of Lists	

A Closer Look at Tier 1

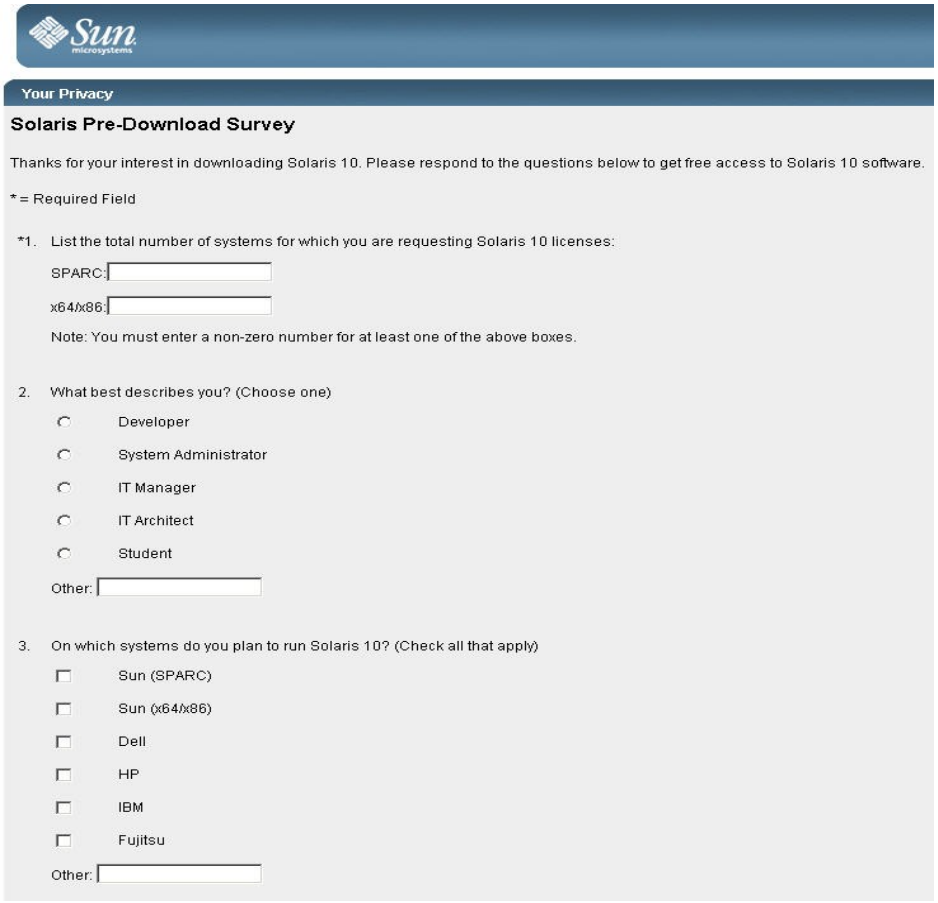
- Users that put in a corporate domain tend to purchase. Users that use free web-based email such as Yahoo or Gmail do not
- People that enter a valid company name tend to purchase
- Opting in for most lists has a positive association with purchase, while a few lists and not opting in at all have a negative association
  - Some lists are better than others. Developers and students do not purchase. SysAdmins and generalists do when they are subscribed to multiple lists

A Closer Look at Tier 2

- Organization (Department)
  - IT, MIS, SYS Admin positions have a high tendency to purchase
- Job Role (more of a functional attribute)
  - The more specific the role that is entered the higher the association with purchase, i.e. Systems Administrator
- The more email lists an individual signs up for increases the association with purchase

Digital Persona Cluster Model

Clustering Data Source – 1 of 2



Your Privacy

Solaris Pre-Download Survey

Thanks for your interest in downloading Solaris 10. Please respond to the questions below to get free access to Solaris 10 software.

* = Required Field

*1. List the total number of systems for which you are requesting Solaris 10 licenses:

SPARC:

x64/x86:

Note: You must enter a non-zero number for at least one of the above boxes.

2. What best describes you? (Choose one)

- Developer
- System Administrator
- IT Manager
- IT Architect
- Student

Other:

3. On which systems do you plan to run Solaris 10? (Check all that apply)

- Sun (SPARC)
- Sun (x64/x86)
- Dell
- HP
- IBM
- Fujitsu

Other:

- Surveys from 20k respondents
  - All within same time frame
  - All requesting Solaris download
- This survey started after modeling data pull
  - Totally independent data used here!
- Question 1 free form inputs coded as 1 for a reasonable answer, 0 for none or silly.
- Rest code 1 if ticked, else 0.
  - “Other” coded 1 if used

Clustering Data Source – 2 of 2

4. Why are you most interested in using Solaris 10? (Check all that apply)

- ☐ Multiplatform support
- ☐ 64-bit support
- ☐ Open Source
- ☐ Solaris Containers (including Zones)
- ☐ Predictable Lifecycle
- ☐ Performance
- ☐ Pricing
- ☐ DTrace
- ☐ Support & Services Offering
- ☐ ZFS

Other: [ ]

5. What type of applications are you working on or using? (Check all that apply)

- ☐ Web Serving (e.g. Apache)
- ☐ Infrastructure Services (e.g. Identity Mgmt, Portal, etc.)
- ☐ Collaboration Services (e.g. Mail, File & Print, etc.)
- ☐ Database Management
- ☐ J2EE Application Services
- ☐ Desktop/Laptop Productivity
- ☐ Platform for Application Development
- ☐ High Performance/Technical Computing

Other: [ ]

Submit and Continue

Variables:

Q1. Licenses?

Q2. Roles?

Q3. Hardware systems?

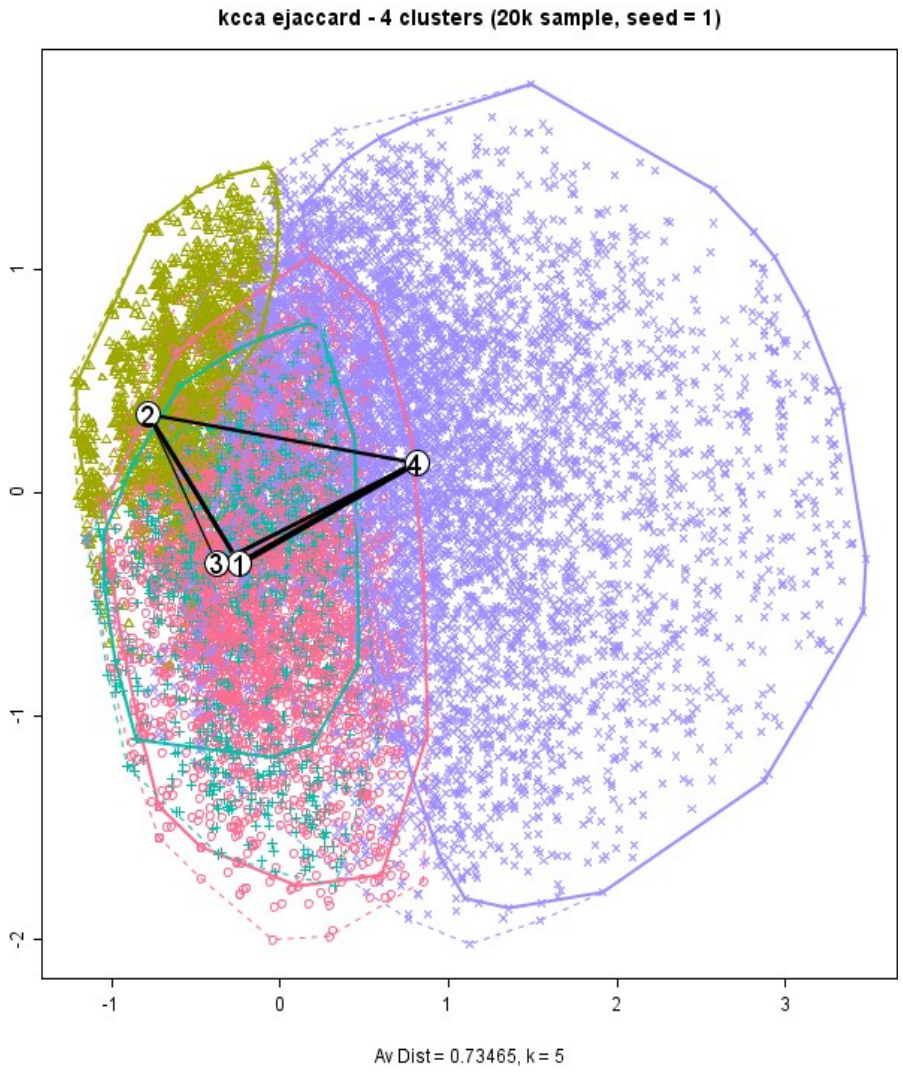
Q4. Why most interested?

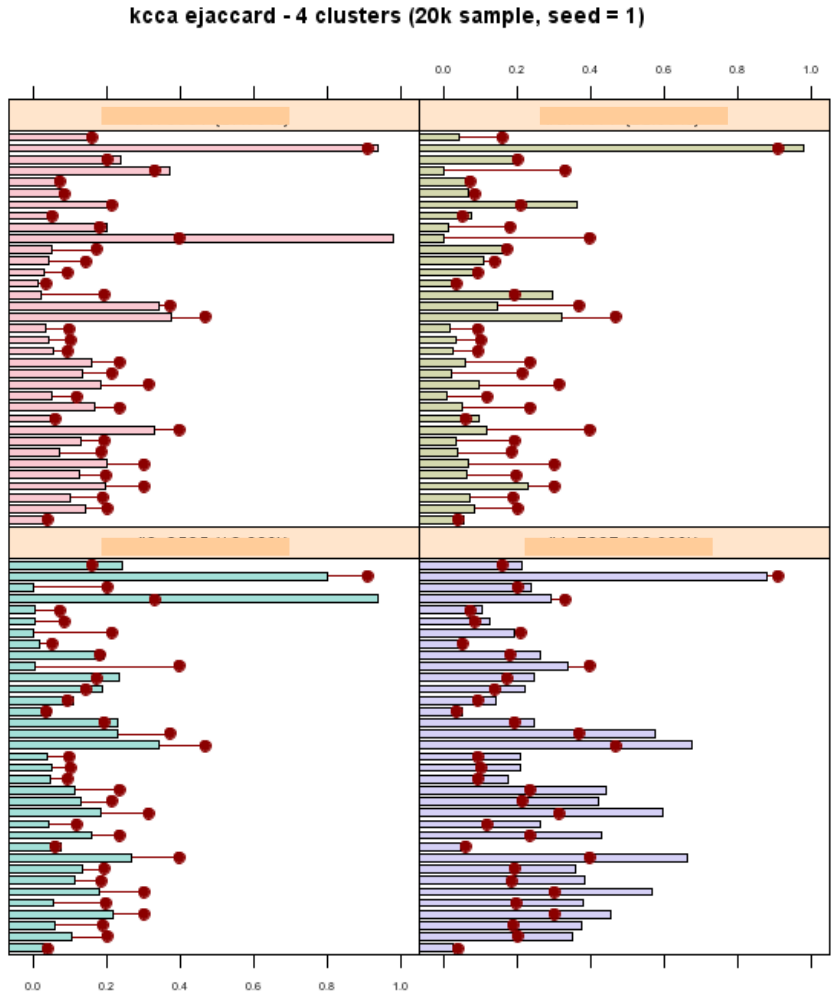
Q5. What applications?

Choice of Clustering Method

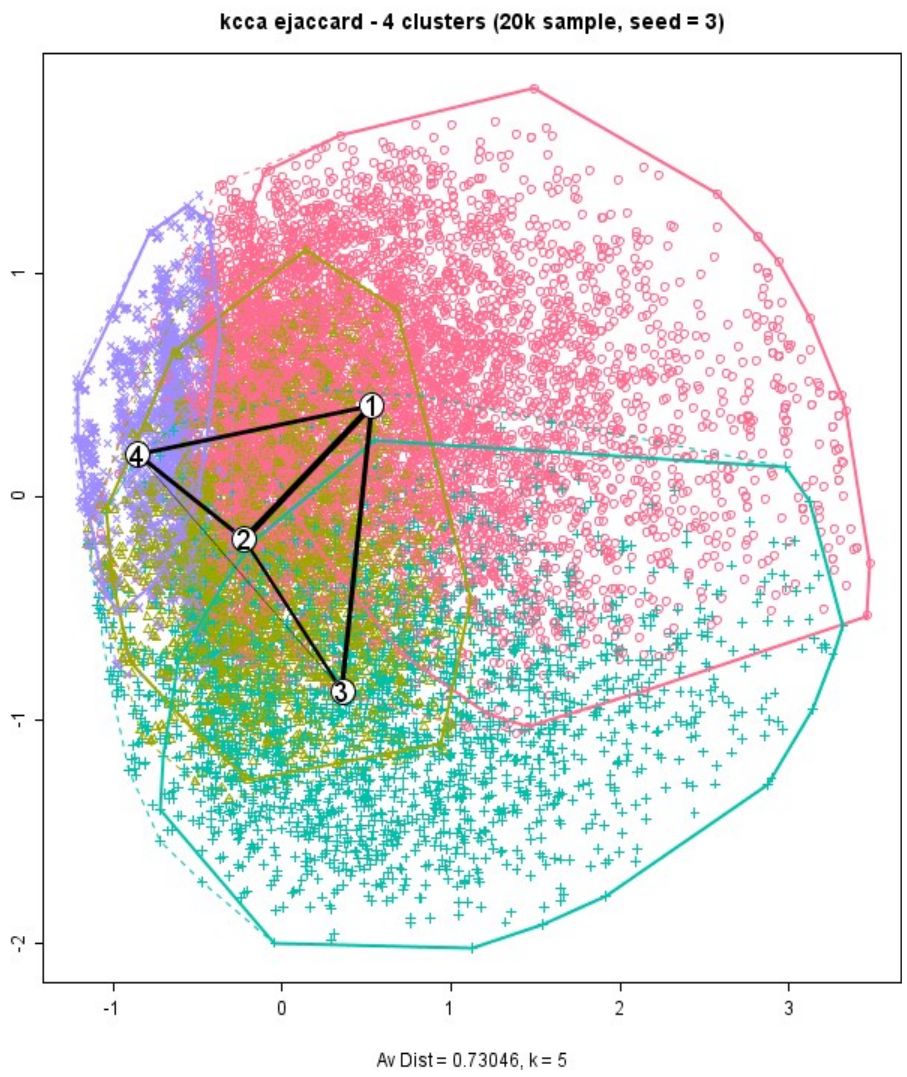
- In this kind of “check box” survey, checking a box is much more significant than not checking it.
  - If I check “Database,” that says a whole lot more about me than if I don't check “Database.”
- We need an appropriate distance measure.
  - Expectation based Jaccard
- Fritz Leisch's flexclust package for R handles this quite nicely.
  - See Appendix for links & details
- We tell flexclust the number of clusters to find. Starting from a random point, it does so.
- Our challenge is picking a stable & meaningful solution.

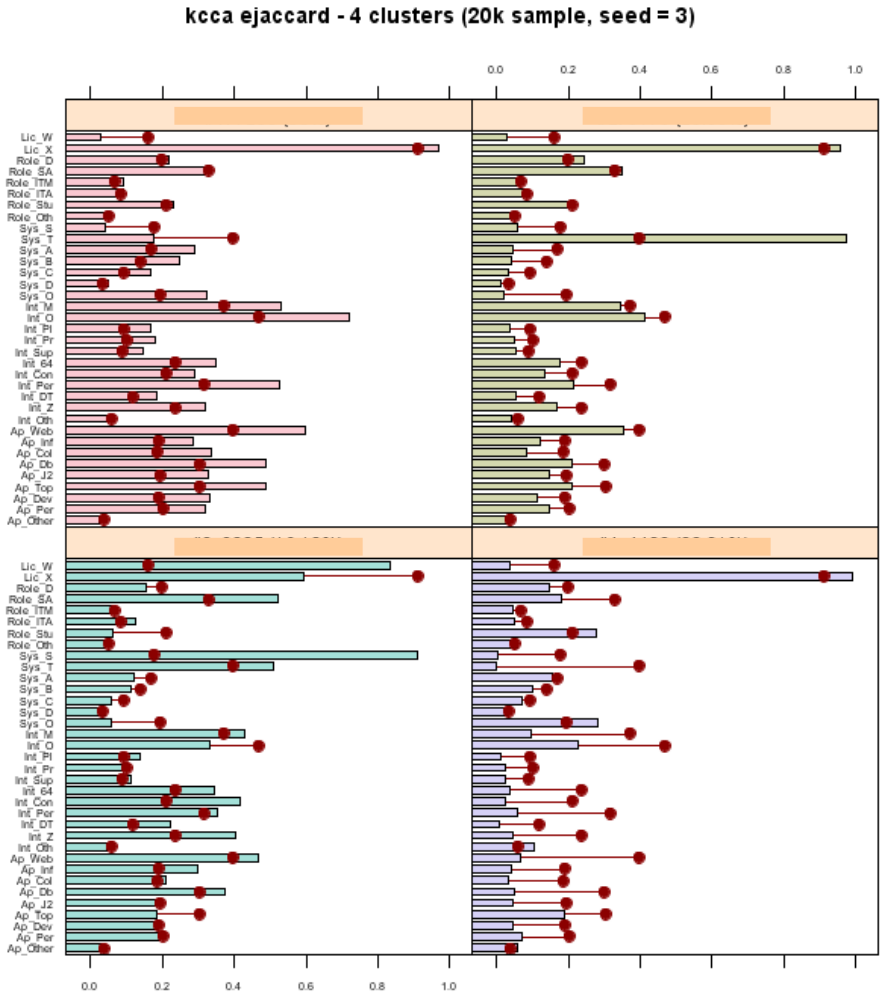
One Example 4-cluster Solution





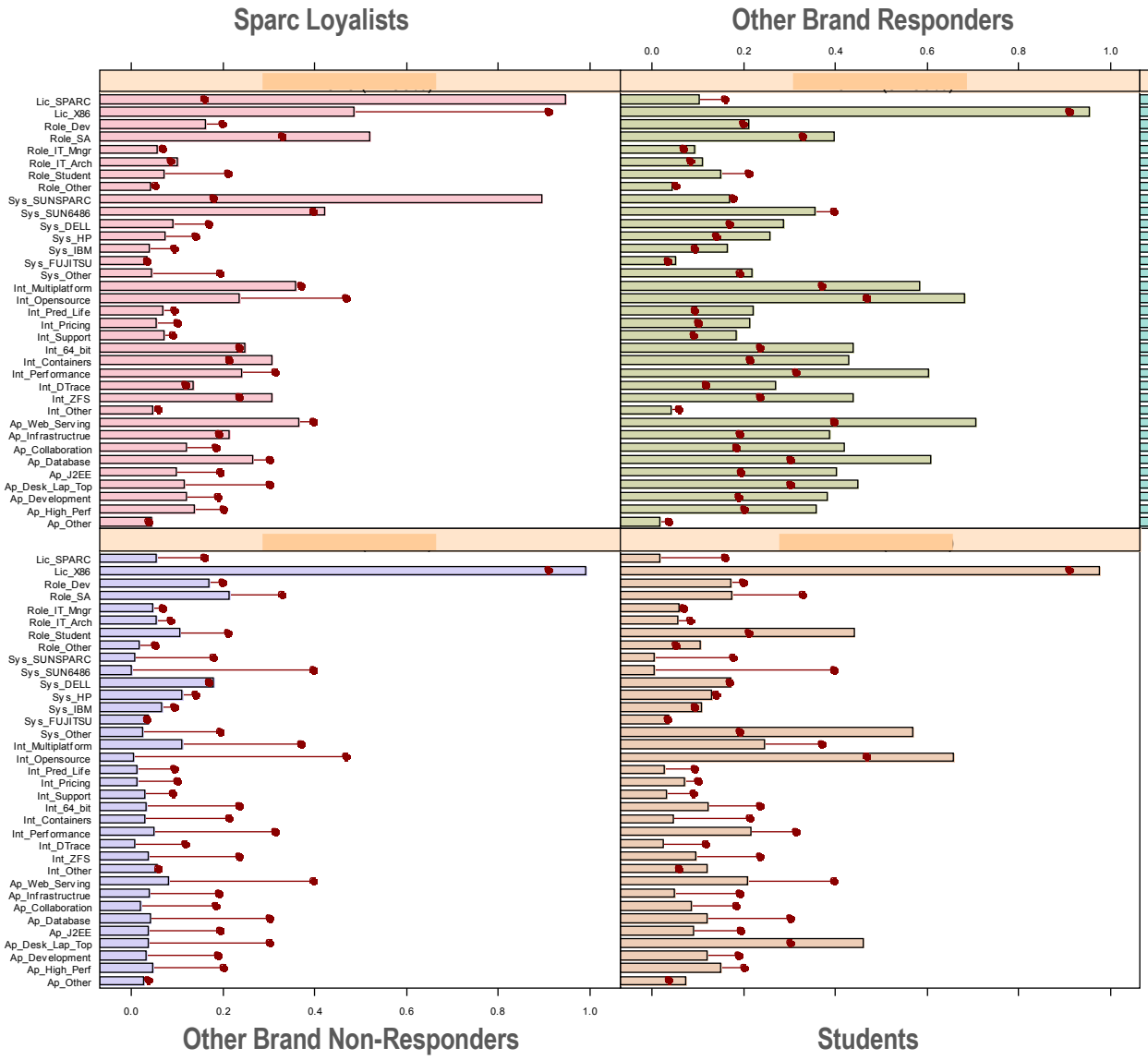
Another Example 4-cluster Solution





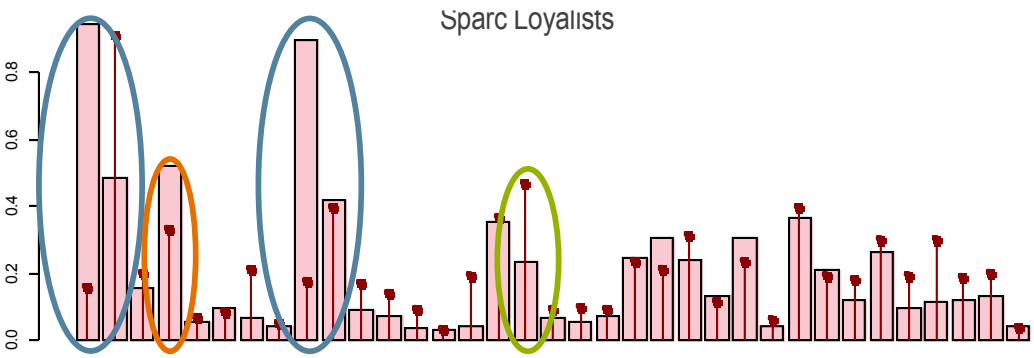
Best Solution was 5 Clusters

- SPARC Loyalists
- Other Brand Responders
- Other Brand Non-Responders
- Sun x86 Loyalists
- Students



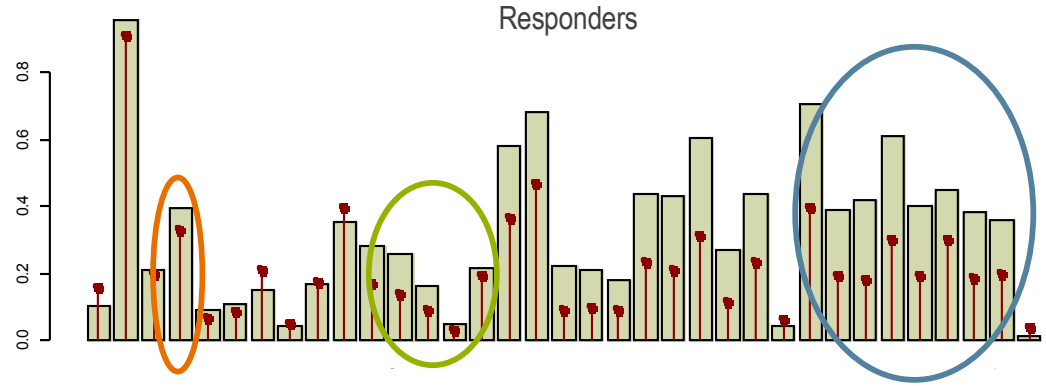
- Red dots represent the average response for the entire population
- Height of the bar represents that clusters response and can be viewed in context of the average response

24



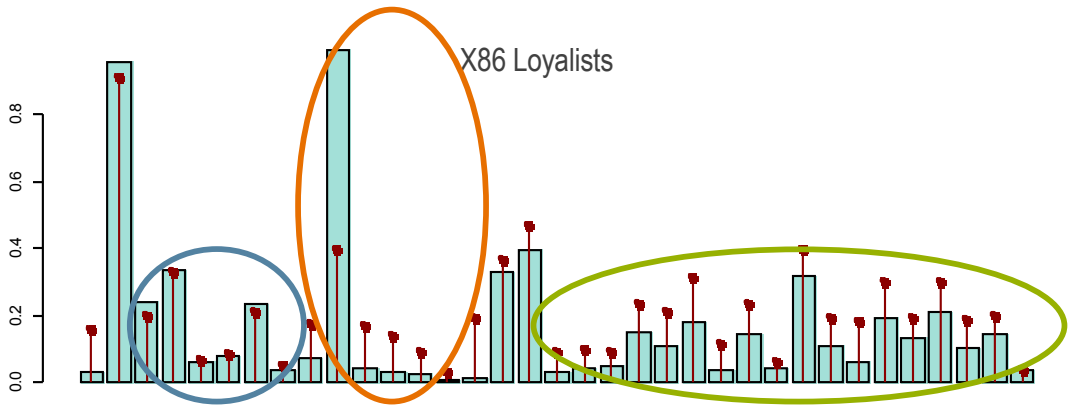
Cluster 1: SPARC Loyalists

- Heavy SPARC users/licensees
- Tend to be Sys Admins
- Do not show interest in open source



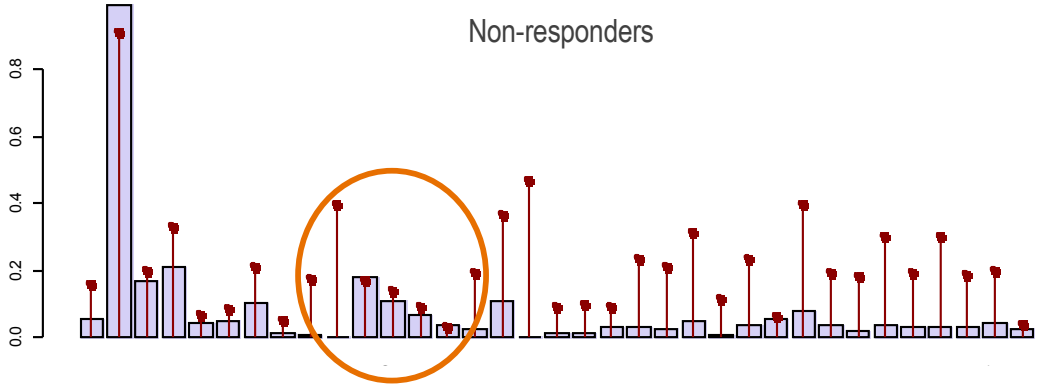
Cluster 2: Other Brand Responders

- Sys admins
- Use multi-platform(non-Sun) hardware in their data centers
- Very Interested in a variety of applications for Solaris 10



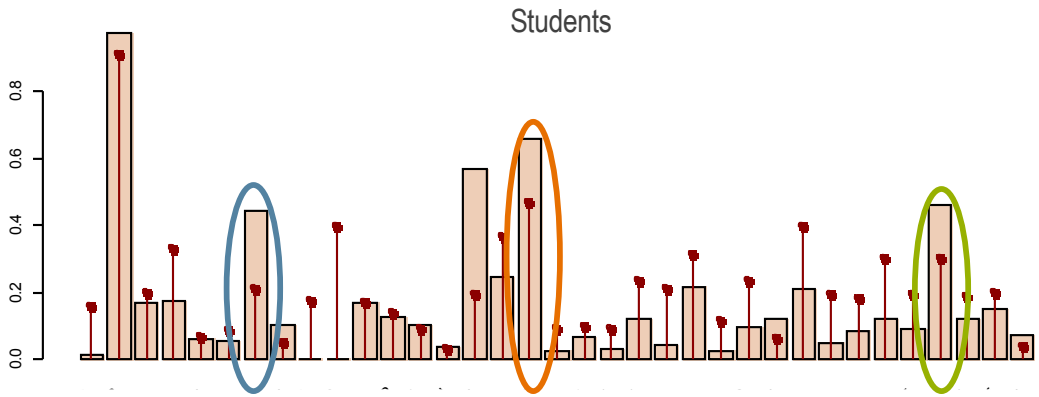
Cluster 3 x86 Loyalists

- Role is not distinguished from the general population
- Plan to run Solaris 10 on Sun x86 hardware. Not on competitor hardware
- Less apt to respond to interest and application questions



Cluster 4 Other Brand Non-responders

- Less likely to respond to questions across the board
- More likely to run Solaris on competitor hardware



Cluster 5 Students

- Identify as students
- Interested in open source
- Desk/Laptop productivity is application of choice

Cluster	Number of Contacts	Number of Purchases	Percent of contacts	Percent Purchases
SPARC Loyalists			Low	High
Other Brand Responders			High	High
X86 Loyalists			High	High
Other Brand Non-Responders			Low	Low
Students			Low	Low

- SPARC loyalists are a small % of x86 downloaders but have a high propensity to purchase
- People who identify as loyal to other hardware providers and responded to the questions are a high % of downloaders and are just as likely to purchase as those that self-identify as loyal Sun x86 hardware
- Students and people that do not fill out the survey completely do not have a high propensity to purchase

Summary of Findings

- Specificity of certain collected registration field values and subscription to multiple specific email lists are highly predictive of purchase propensity (Tier 1 & Tier 2)
- Incompleteness of registration response is predictive of low purchase propensity
- Stated preferences for competitor products does not negatively impact purchase propensity
- Students have more time than money

Shifts in Marketing Strategy

- Help purchasers self-identify
- Focus on registration standardization
- Targeted content development
- Segmented outbound calling lists
- Incorporate learnings into our lead scoring algorithm – in progress

Early Results

- 2x increase in ratio of leads to outbound calls
- 33% increase in per-lead pipeline $ value
- 3 – 4x improvement in outbound calling efficiency
- FY '09 goal for standardized registration

Questions?

- Now would be the time!
- Keep in contact!
  - Alex's email: alex.kriney@sun.com
  - Jim's email: JPorzak@tgn.com

Appendix

Sun Microsystems Links

Product/Solution	URL
OpenStorage	sun.com/storagetek/open.jsp
Sun Systems for MySQL	sun.com/systems/solutions/mysql/
Sun x64 Systems	sun.com/x64/index.jsp
Java	java.com
NetBeans	netbeans.org
OpenSolaris	opensolaris.org
MySQL	mysql.com
GlassFish	sun.com/glassfish
OpenSSO	opensso.dev.java.net/
JavaFX	javafx.com
OpenOffice	openoffice.org
xVM VirtualBox	virtualbox.org
OpenSPARC	opensparc.org
Solaris	sun.com/software/solaris/

R Links

- R Homepage: www.r-project.org
  - The official site of R
- R Foundation: www.r-project.org/foundation
  - Central reference point for R development community
  - Holds copyright of R software and documentation
- Local CRAN mirror site:
  - We use: cran.cnr.berkeley.edu
  - Find yours at: cran.r-project.org/mirrors.html
  - Current Binaries, Documentation & FAQs, & more!
- Your local useR! Group:
  - www.meetup.com/R-Users/

Random Forest Links

- Original Fortran 77 source code freely available from Breiman & Cutler:
  - http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_h
    &
    http://www.math.usu.edu/~adele/forests/
- Commercialization by Salford Systems:
  - http://www.salford-systems.com/randomforests.php
- R package, randomForest. An adaptation by Andy Liaw & Matthew Wiener of Merck:
  - http://cran.cnr.berkeley.edu/web/packages/randomForest/ind
- See Jim's randomForest tutorial deck:
  - http://www.porzak.com/JimArchive/JimPorzak_RFwithR_DMAAC_Jan07_

flexclust Links

- The flexclust package on CRAN:
  - http://cran.cnr.berkeley.edu/web/packages/flexclust/index.html
- Fritz Leisch's page:
  - http://www.stat.uni-muenchen.de/~leisch/
    - (despite picture, he is really a jolly guy!)
  - See: http://www.stat.uni-muenchen.de/~leisch/papers/Leisch-2006.pdf
- See second half of Jim's useR! 2008 tutorial:
  - http://www.porzak.com/JimArchive/useR2008/useR08_JimP_Cus