DMA Direct Marketing Association *Analytics Council Virtual Seminar:*

Predictive Modeling with Random Forests™ in R

A Practical Introduction to R for Business Analysts

by Jim Porzak

January 2007

Loyalty Matrix, Inc.
580 Market Street, 6th Floor
San Francisco, CA 94104
(415) 296-1141
www.loyaltymatrix.com

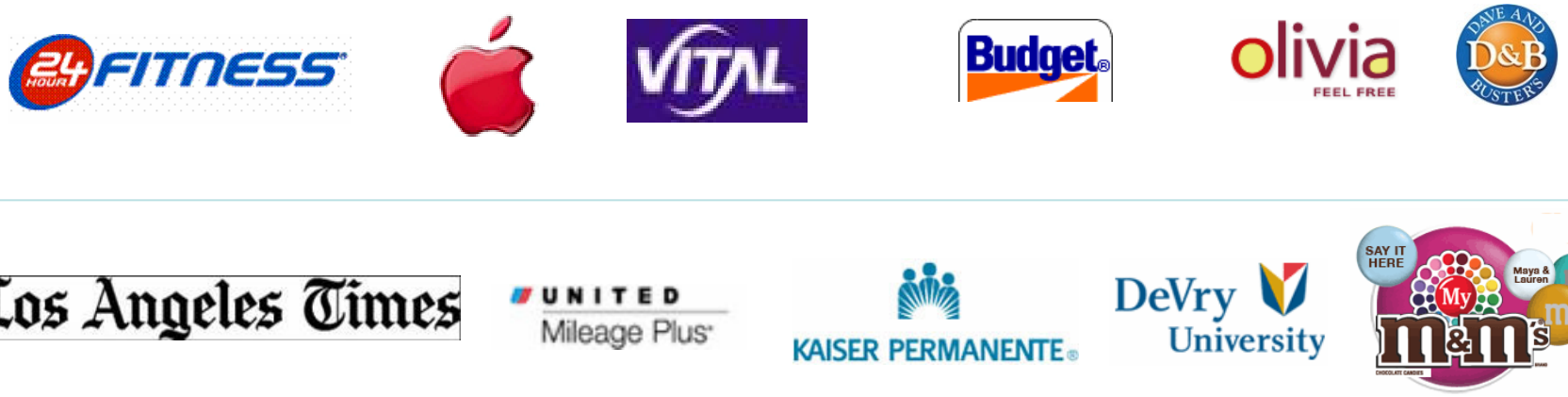
Outline

- Part I – Introduction to R
- Part II – Using Random Forests for Classificaiton
- Wrap up & Questions/Discussion

Note: For R setup details see first Appendix slide.

Background on Loyalty Matrix

- Provide customer data analytics to optimize direct marketing resources
- OnDemand platform MatrixOptimizer® (version 3.2)
- Over 20 engagements with Fortune 500 clients
- Experienced team with diverse skills & backgrounds
- 10-person San Francisco firm with an offshore team in Nepal



Part I – Introduction to R

- Background
  - Evolution & History
  - Current state of R
  - Resources
    - R-Help
    - Task Views
- Simple code example
  - Functions
  - Objects
  - Methods
- Where to learn more

Evolution of R from 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
  - "S-Plus" now owned by Insightful Corp., Seattle, WA
- 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 in "HAL" is before "IBM"
- Since 1997 a core group of ± 20 developers
  - Initial V1.0 released in February, 2000
  - Continually developed with a new 0.1 level release ~ 6 months

Current state of R

- V2.4.1 Released December, 2006
- Windows, Mac OS, Linux & Unix ports
- Now 931 submitted packages from “aaMI” to “zoo”
- 18th newsletter (Volume 6/5) published December 2006
- The second useR! conference– Vienna June 2006
- Dozens of texts specifically on R or using R examples
- R language generally accepted to be more powerful than S-Plus
- Some interesting GUI work in progress - JGR

R Resources

- 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
    - I use: cran.cnr.berkeley.edu/
    - Find your's at: cran.r-project.org/mirrors.html
  - Current Binaries
  - Current Documentation & FAQs
  - Links to related projects and sites
- Mailing Lists
  - Best help ever!

R-Help Mailing List Example

Core Developers!

Remove label "R-Help" | Report Spam | Delete | More actions... | Refresh | 1 - 50 of 10513 Older › Oldest »

Select: All, None, Read, Unread, Starred, Unstarred

march	Inbox **[R] gbn with jumps** - Hi everybody I'd like to simulate a Generalized Wiener Process with jumps. Any sugge	7:03 am
Vladimir Eremeev	Inbox **[R] simpler solution (untested)** - axis says that this function has the logical parameter outer "indicating whe	6:56 am
march	Inbox **[R] gbm with jumps** - Hi everybody I'd like to simulate a Generalized Wiener Process with jumps. Any sugge	6:48 am
Rafael, **Peter**, **Vladimir** (3)	Inbox **[R] Three horizontal axes OR Two axes on same side?** - Dear list: I need to reproduce a plot with three d	6:44 am
Bram Kuijper	Inbox **[R] levelplot not adjusting colors** - Hi all, I try to make a levelplot from the Trellis graphics package of count	6:41 am
Marta Rufino	Inbox **[R] warning in GAM** - Hello, I have a problem when doing gam (from gam library; I am using R 2.4.0, windows	5:48 am
Antje, **Peter** (4)	Inbox **[R] Error in plot.new() : Figure margins too large** - Hello, was could be the reason for such an error messa	5:33 am
Indermaur, **Ken**, **Prof** (3)	Inbox **[R] batch job GLM calculations** - Hello I want to batch job the calculation of many GLM-models, extract sor	1:19 am
Adrian .. **Prof**, **Adrian** (9)	Inbox **[R] a question of substitute** - The 'Right Thing' is for oneway.test() to allow a variable for the first argument, a	12:42 am
David, **Marc** (2)	Inbox **[R] zero margin / marginless plots** - Hi, I'd like to produce a marginless or zero margin plot so that the pixe	7:37 pm
Walter, **Torsten**, **Richard** (3)	Inbox **[R] posthoc tests with ANCOVA** - The WoodEnergy example in package HH (available on CRAN) is similar #	Jan 10
karl.sommer	Inbox **[R] axis date format in lattice** - Hello list, plotting the following example 1 in lattice only labels the x-axis wi	Jan 10
Tong .. Prof, **François** (9)	Inbox **[R] A question about R environment** - Philippe Grosjean] >Please, don't reinvent the wheel: putting functior	Jan 10
Michael, **Peter** (2)	Inbox **[R] TCL/TK and R documentation?** - I am hoping something has changed since I last asked about this. Is tl	Jan 10
Simon, Setzer.Wood., **Ken** (3)	Inbox **[R] problems with optim, "for"-loops and machine precision** - Two possibilities for why your 7 parameter	Jan 10
Darren Weber	Inbox **[R] axis labels at subset of tick marks** - For example, this works: x = seq(-100, 1000, 25) y = x * x plot(x,y,	Jan 10
Colleen.Ross .. **Thomas** (3)	Inbox [R] SAS and R code hazard ratios - On Wed, 10 Jan 2007, Colleen.Ross@kp.org wrote: > I am new to R and	Jan 10
Thomas, **Duncan**, Peter (3)	Inbox [R] "go" or "goto" command - Thomas L Jones wrote: > Some computer languages, including C, have a "go" c	Jan 10
Feng, David, **Feng** (3)	Inbox **[R] logistic regression packages** - Hi David: Thanks for you information. 2 further questions: 1. I found out th	Jan 10
David	Inbox [R] Installation problem with package mixtools - I am trying to install mixtools on Debian Etch and get the follc	Jan 10
Tord, **Roger** (2)	Inbox [R] map data.frame() data after having linked them to a read.shape() object - On Wed, 10 Jan 2007, Tord Snäl	Jan 10
Stephen, chao (2)	Inbox [R] using DBI - The way MySQL works, I use RMySQL to contact, which in turn uses DBI. There is a library R	Jan 10
Paul Mathews	Inbox [R] Meeting announcement: An Introduction to Data Analysis Using R - An Introduction to Data Analysis Usin	Jan 10
Kati, roger (2)	Inbox [R] 2 problems with latex.table (quantreg package) - reproducible - The usual R-help etiquette recommends: 1	Jan 10
John .. Jeffrey, Brian (12)	Inbox [R] scripts with littler - Brian Ripley wrote: > Exactly as documented. The argument is named 'new' and not ...	Jan 10
Jenny, Zoltan (3)	Inbox [R] correlation value and map - Hi Zoltan, Right, I have 30x32=960 data points per year (It is actually the mear	Jan 10

CRAN Task Views

Quick start guides to packages by task at hand

Bayesian	Bayesian Inference
Cluster	Cluster Analysis & Finite Mixture Models
Econometrics	Computational Econometrics
Environmetrics	Analysis of ecological and environmental data
Finance	Empirical Finance
Genetics	Statistical Genetics
Graphics	**Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization**
MachineLearning	**Machine Learning & Statistical Learning**
Multivariate	Multivariate Statistics
SocialSciences	Statistics for the Social Sciences
Spatial	Analysis of Spatial Data
gR	gRaphical models in R

Link: cran.cnr.berkeley.edu/src/contrib/Views/

R Basics – Functions

- A useful function for DMers…

```
> prop.test(c(138, 113), c(2500, 2500))
```

```
2-sample test for equality of proportions with continuity correction

data:  c(138, 113) out of c(2500, 2500)
X-squared = 2.4161, df = 1, p-value = 0.1201
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.002501721  0.022501721
sample estimates:
prop 1 prop 2
0.0552 0.0452
```

R Basics - Objects

- Most functions return an object

```
pt <- prop.test(c(138, 113), c(2500, 2500))
> str(pt)
```

```
List of 9
 $ statistic  : Named num 2.42
  ..- attr(*, "names")= chr "X-squared"
 $ parameter  : Named num 1
  ..- attr(*, "names")= chr "df"
 $ p.value    : num 0.12
 $ estimate   : Named num [1:2] 0.0552 0.0452
  ..- attr(*, "names")= chr [1:2] "prop 1" "prop 2"
 $ null.value : NULL
 $ conf.int   : atomic [1:2] -0.0025  0.0225
  ..- attr(*, "conf.level")= num 0.95
 $ alternative: chr "two.sided"
 $ method     : chr "2-sample test for equality of proportions with
continuity correction"
 $ data.name  : chr "c(138, 113) out of c(2500, 2500)"
 - attr(*, "class")= chr "htest"
```

R Basics - Methods

- Objects have methods…
- One of which we have used already – the default

```
print(pt)
```

```
2-sample test for equality of proportions with continuity correction

data:  c(138, 113) out of c(2500, 2500)
X-squared = 2.4161, df = 1, p-value = 0.1201
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.002501721  0.022501721
sample estimates:
prop 1 prop 2
0.0552 0.0452
```

R Basics – Learning More

- Wikipedia
  - http://en.wikipedia.org/wiki/R_%28programming_language%29
- *An Introduction to R*
  - http://cran.cnr.berkeley.edu/doc/manuals/R-intro.html
- Links to all "official" manuals (htlm & 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

Part II – Random Forests

- Background
  - History
  - Advantages
  - Versions
- Example walkthrough using R
  - Problem & Data Descriptions
  - Data Prep
  - Building the Forest
  - Diagnostics
  - Interpretation
  - Prediction
  - Scoring
- And More…

Random Forests - History

- Developed by Leo Breiman of Cal Berkeley, one of the four developers of CART, and Adele Cutler, now at Utah State University.
- An extension of single decision tree methods like CART & CHAID.
- Many small trees are randomly grown to build the forest. All are used in the final result.
- See Wikipedia for more
  - http://en.wikipedia.org/wiki/Random_forest

Random Forests - Advantages

- 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)

Random Forests - Versions

- Original Fortran 77 source code freely available from Breiman & Cutler.
  http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm
  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 of Merck.
  http://cran.cnr.berkeley.edu/src/contrib/Descriptions/randomForest.html

RF Example – Description

- Sample Data from a sports club
- Challenge – predict “at-risk” members based on membership usage data & simple demographics
- Training & Test data sets provided:
  - MemberTrainingSet.txt (1916 records)
  - MemberTestSet.txt (1901 records)

RF Example – Columns

- MembID (identifier)
- Status = M or C
  (Member or Cancel)
- Gender
- Age
- MembDays
- NumUses1st30d
- NumUsesLast30d
- TotalUses
- FirstCkInDay
- LastCkInDay
- DaysSinceLastUse
- TotalPaid
- MonthlyAmt
- MilesToClub
- NumExtras1st30d
- NumExtrasLast30d
- TotalExtras
- DaysSinceLastExtra

RF Example – Getting Started

- Load the randomForest package

```
> ## CIwR_rf.R
> require(randomForest)
```

```
Loading required package: randomForest
randomForest 4.5-18
Type rfNews() to see new features/changes/bug fixes.
[1] TRUE
```

- Point to working environment

```
> setwd("c:/Projects/CIwR/R")
> dir("Data")
```

```
[1] "CruiseReservationEvents.txt" "KeyCustomers.txt"
[3] "MemberTestSet.txt"           "MemberTrainingSet.txt"
[5] "NewSubscribers.txt"          "orders.txt"
[7] "ZipPopDist.txt"
```

RF Example – Load Training Data

```
> Members <- read.delim("Data/MemberTrainingSet.txt", row.names = "MembID")
> str(Members)
```

```
`data.frame':   1916 obs. of  17 variables:
 $ Status             : Factor w/ 2 levels "C","M": 1 1 1 1 1 1 1 1 1 1 ...
 $ Gender             : Factor w/ 3 levels "F","M","U": 2 2 1 2 2 1 1 2 ...
 $ Age                : int  21 18 21 21 45 25 21 20 35 15 ...
 $ MembDays           : int  92 98 30 92 31 249 1 92 322 237 ...
 $ NumUses1st30d      : int  11 11 3 6 24 2 0 16 12 6 ...
 $ NumUsesLast30d     : int  6 6 3 1 24 0 0 4 0 0 ...
 $ TotalUses          : int  28 31 3 9 24 6 0 30 38 26 ...
 $ FirstCkInDay       : Factor w/ 556 levels "","2004-01-04",..: 132 264 ...
 $ LastCkInDay        : Factor w/ 489 levels "","2004-01-15",..: 134 356 ...
 $ DaysSinceLastUse   : int  3 2 9 11 4 196 NA 12 138 65 ...
 $ TotalPaid          : int  149 136 100 129 75 134 138 149 582 168 ...
 $ MonthlyAmt         : int  NA 27 NA NA NA 31 30 NA NA 10 ...
 $ MilesToClub        : int  4 0 0 5 2593 4 5 4 NA 2 ...
 $ NumExtras1st30d    : int  0 0 0 0 0 0 0 0 1 0 ...
 $ NumExtrasLast30d   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ TotalExtras        : int  0 0 0 0 0 0 0 0 6 0 ...
 $ DaysSinceLastExtra: int  NA NA NA NA NA NA NA NA 253 NA ...
```

RF Example – Quick Look at Data (1 of 2)

```
> summary(Members)
```

```
 Status    Gender        Age             MembDays      NumUses1st30d
 C: 809    F:870    Min.   :13.00    Min.   :  1.0    Min.   : 0.000
 M:1107    M:832    1st Qu.:23.00    1st Qu.: 92.0    1st Qu.: 1.000
           U:214    Median :29.00    Median :220.0    Median : 4.000
                    Mean   :32.72    Mean   :247.8    Mean   : 5.385
                    3rd Qu.:40.00    3rd Qu.:365.0    3rd Qu.: 8.000
                    Max.   :82.00    Max.   :668.0    Max.   :36.000
                    NA's   : 1.00

 NumUsesLast30d     TotalUses          FirstCkInDay      LastCkInDay
 Min.   : 0.000    Min.   :  0.00              : 236              : 236
 1st Qu.: 0.000    1st Qu.:  3.00    2004-06-01:  10    2005-10-28:  56
 Median : 0.000    Median : 12.00    2004-06-23:  10    2005-10-27:  55
 Mean   : 2.125    Mean   : 26.73    2004-11-01:  10    2005-10-30:  52
 3rd Qu.: 3.000    3rd Qu.: 33.00    2005-02-02:  10    2005-10-26:  47
 Max.   :26.000    Max.   :340.00    2004-09-13:   9    2005-10-29:  42
                                     (Other)   :1631    (Other)   :1428
```

Continued on next slide…

RF Example – Quick Look at Data (2 of 2)

```
                                … Continued from above
   DaysSinceLastUse    TotalPaid         MonthlyAmt       MilesToClub
   Min.    :  1.00   Min.    :  0.00   Min.    :  4.00   Min.    :   0.00
   1st Qu.:  7.00   1st Qu.: 70.75   1st Qu.: 21.00   1st Qu.:   1.00
   Median : 32.00   Median :135.00   Median : 28.00   Median :   3.00
   Mean    : 75.51   Mean    :188.75   Mean    : 28.50   Mean    :  24.40
   3rd Qu.:106.00   3rd Qu.:232.25   3rd Qu.: 35.00   3rd Qu.:   7.00
   Max.    :624.00   Max.    :961.00   Max.    : 94.00   Max.    :2609.00
   NA's    :236.00                     NA's    :536.00   NA's    : 202.00

 NumExtras1st30d   NumExtrasLast30d    TotalExtras      DaysSinceLastExtra
 Min.    : 0.0000   Min.    : 0.00000   Min.    :  0.000   Min.    :   2.00
 1st Qu.: 0.0000   1st Qu.: 0.00000   1st Qu.:  0.000   1st Qu.:  55.25
 Median : 0.0000   Median : 0.00000   Median :  0.000   Median : 195.00
 Mean    : 0.4128   Mean    : 0.09603   Mean    :  1.324   Mean    : 229.85
 3rd Qu.: 0.0000   3rd Qu.: 0.00000   3rd Qu.:  0.000   3rd Qu.: 376.00
 Max.    :13.0000   Max.    :14.00000   Max.    :121.000   Max.    : 660.00
                                                          NA's    :1646.00
```

- Absolute Dates not useful (at least down to day level)
- RF does not like NA's!
  - Day's Since Last xxx is NA when no event, use large # days
  - Impute remaining NA's

RF Example – Prepping the data set

- Subset out the absolute dates:

```
> Members <- subset(Members, select = -c(FirstCkInDay, LastCkInDay))
```

- Replace days since last NA's with 999:

```
> Members$DaysSinceLastUse[is.na(Members$DaysSinceLastUse)] <- 999
> Members$DaysSinceLastExtra[is.na(Members$DaysSinceLastExtra)] <- 999
```

- Impute remaining NA's with Random Forests' impute:

```
> Members <- rfImpute(Status ~ ., data = Members)          ## 70 sec
```

```
ntree      OOB      1      2
  300:  21.82% 31.64% 14.63%
ntree      OOB      1      2
  300:  22.44% 33.13% 14.63%
ntree      OOB      1      2
  300:  21.76% 31.89% 14.36%
ntree      OOB      1      2
  300:  21.45% 32.14% 13.64%
ntree      OOB      1      2
  300:  20.72% 31.64% 12.74%
```

RF Example – One Last Look at Data & Save It

```
> summary(Members)
```

```
 Status    Gender        Age             MembDays      NumUses1st30d
 C: 809    F:870    Min.   :13.00    Min.   :  1.0    Min.   : 0.000
 M:1107    M:832    1st Qu.:23.00    1st Qu.: 92.0    1st Qu.: 1.000
           U:214    Median :29.00    Median :220.0    Median : 4.000
                    Mean   :32.71    Mean   :247.8    Mean   : 5.385
                    3rd Qu.:40.00    3rd Qu.:365.0    3rd Qu.: 8.000
                    Max.   :82.00    Max.   :668.0    Max.   :36.000

 NumUsesLast30d     TotalUses       DaysSinceLastUse   TotalPaid
 Min.   : 0.000    Min.   :  0.00    Min.   :  1.0    Min.   :  0.00
 1st Qu.: 0.000    1st Qu.:  3.00    1st Qu.:  9.0    1st Qu.: 70.75
 Median : 0.000    Median : 12.00    Median : 47.0    Median :135.00
 Mean   : 2.125    Mean   : 26.73    Mean   :189.3    Mean   :188.75
 3rd Qu.: 3.000    3rd Qu.: 33.00    3rd Qu.:172.0    3rd Qu.:232.25
 Max.   :26.000    Max.   :340.00    Max.   :999.0    Max.   :961.00

                              … more cut …
```

```
> save(Members, file = "MemberTrainingSetImputed.rda")
```

RF Example – Building the Forest!

```
> Members.rf <- randomForest(Status ~ ., data = Members, importance = TRUE,
proximity = TRUE)                                          ## 30 sec

> Members.rf
```

```
Call:
 randomForest(x = Members[-1], y = Members$Status, ntree = 500,
              mtry = 3, importance = TRUE, proximity = TRUE, data = Members)

               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 21.4%
Confusion matrix:
    C   M class.error
C 546 263   0.3250927
M 147 960   0.1327913
```

- Rather good results. Only ~20% overall error rate.
  - 33% false positive
  - 13% false negative

RF Example – Tuning the Forest

- ntree = 500 & mtry = 3 are defaults. Try tuning them.

```
> Members.rf <- randomForest(Members[-1], Members$Status, data = Members,
mtry = 4, ntree = 1000, importance = TRUE, proximity = TRUE)     ## 50 sec
> Members.rf
```

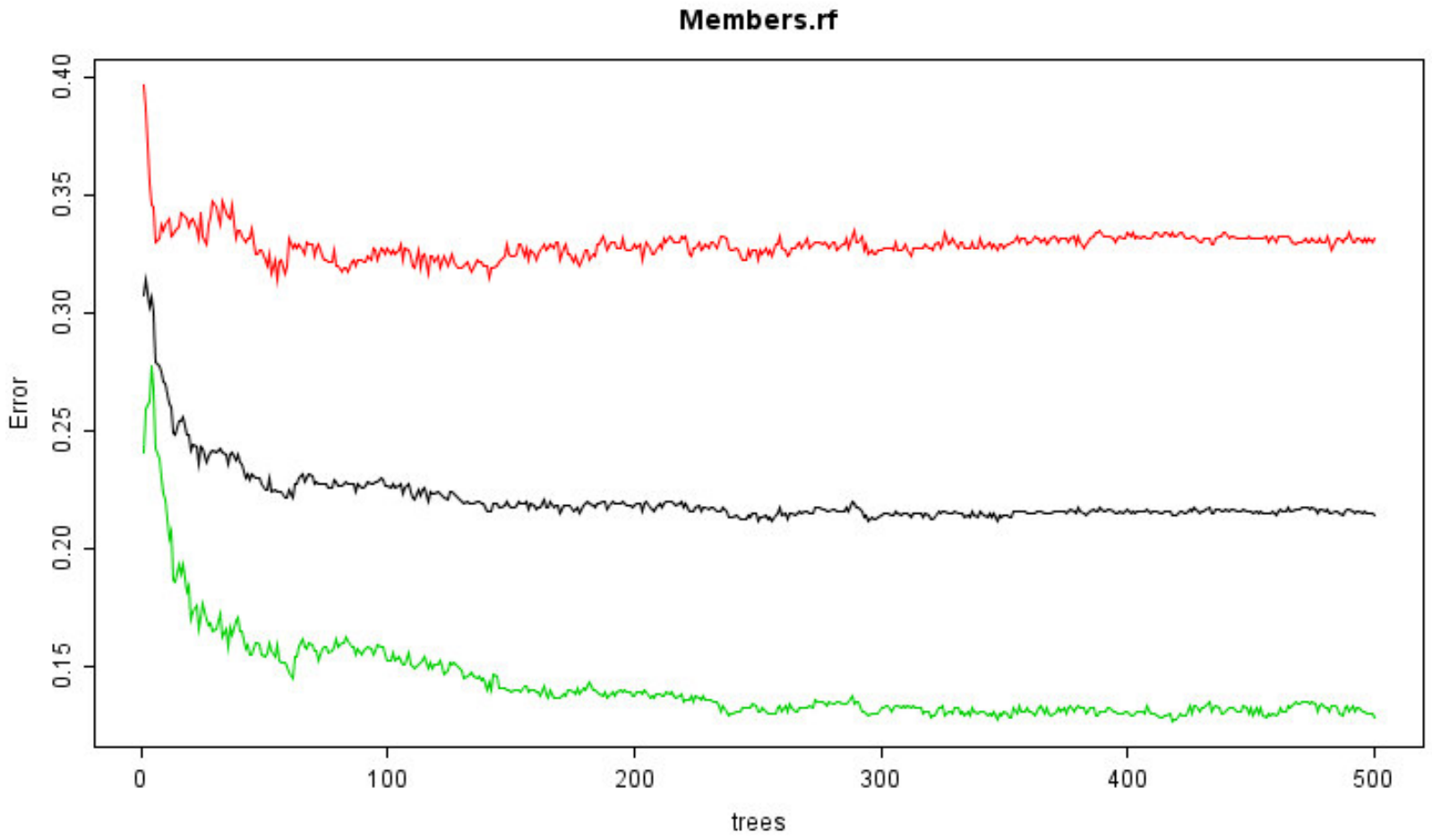
```
Call:
 randomForest(x = Members[-1], y = Members$Status, ntree = 1000, mtry = 4,
importance = TRUE, proximity = TRUE, data = Members)

               Type of random forest: classification
                     Number of trees: 1000
No. of variables tried at each split: 4

        OOB estimate of  error rate: 21.14%
Confusion matrix:
    C   M class.error
C 556 253   0.3127318
M 152 955   0.1373080
```

- No real difference (probably within random effects)

RF Example – Diagnostics (1 of 3)

- RF Diagnostics - OOB errors by # trees

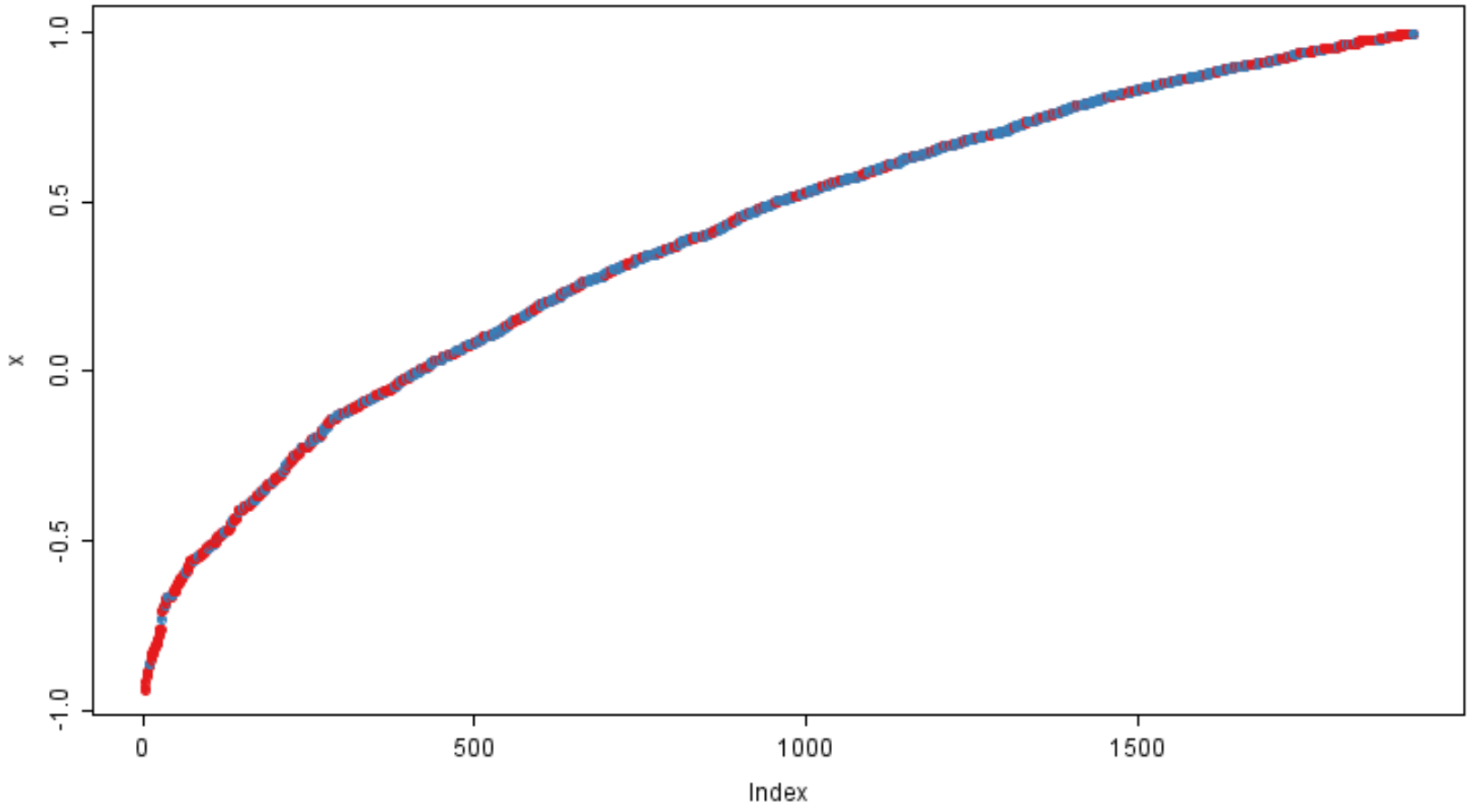
```
> Plot(Members.rf, lty = 1)
```



RF Example – Diagnostics (2 of 3)

- RF Diagnostics – Margin Plot

```
> plot(margin(Members.rf, Members$Status))
```

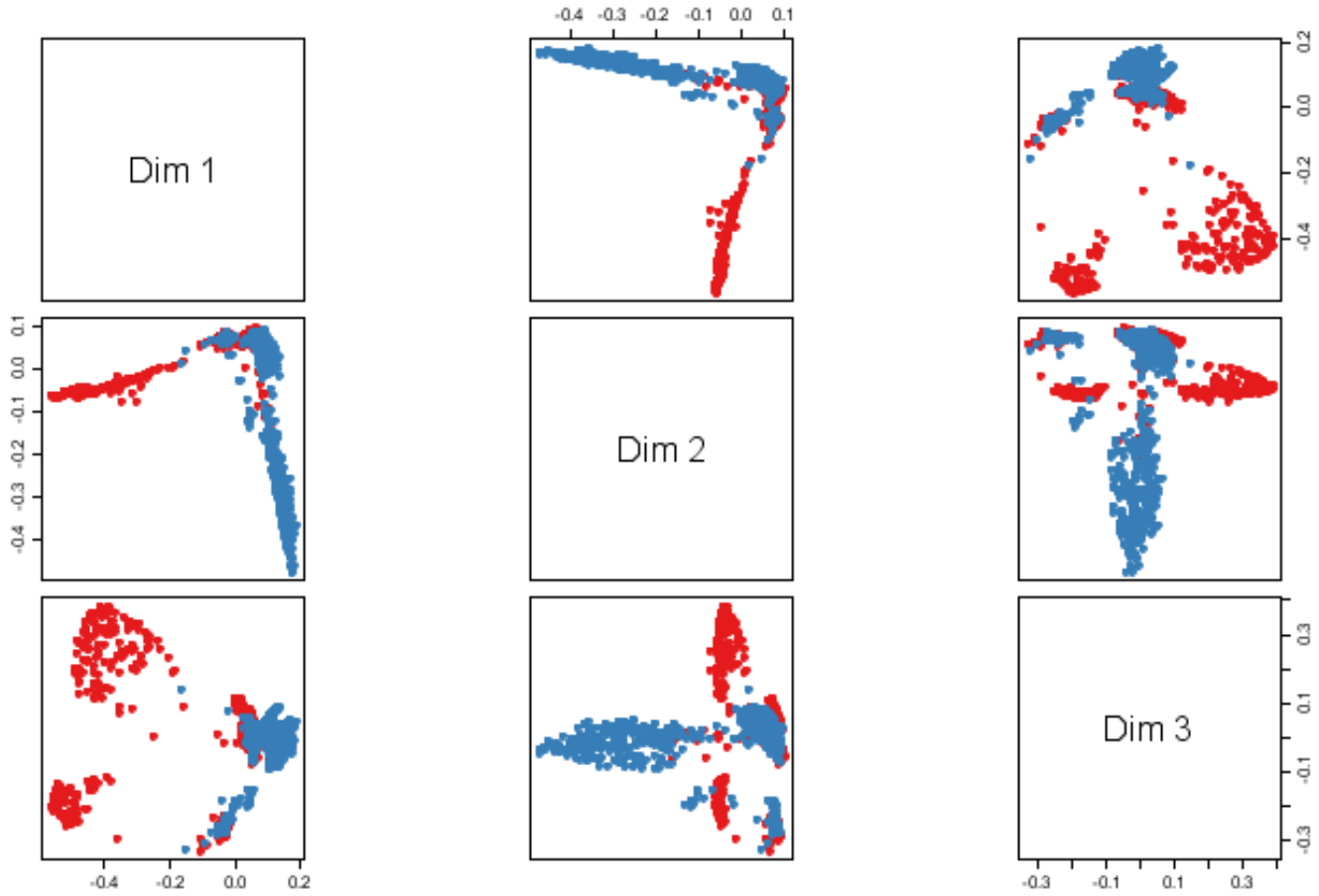


Loyalty Matrix

RF Example – Diagnostics (3 of 3)

- RF Diagnostics – MDS Plot

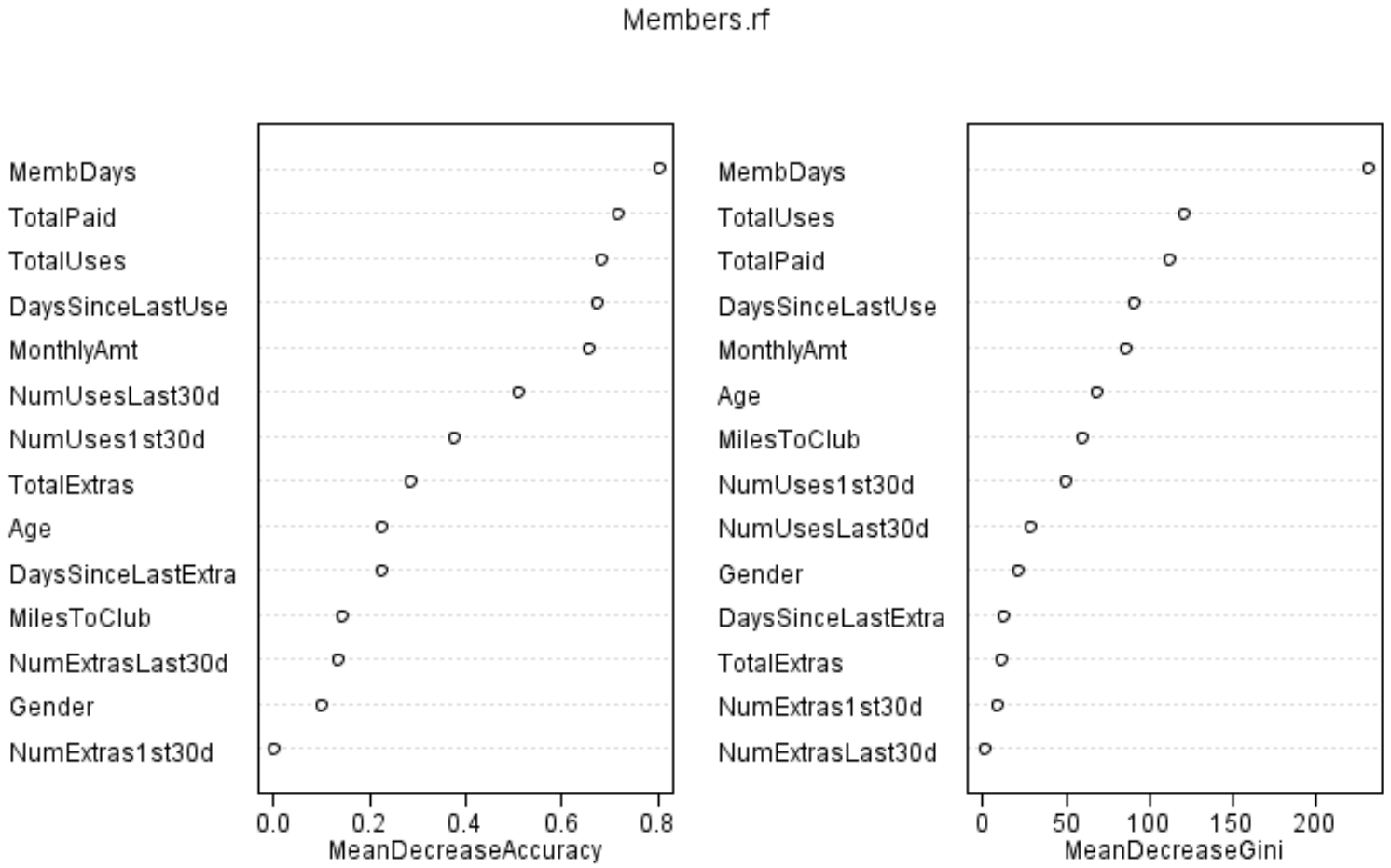
```
> MDSplot(Members.rf, Members$Status, k =3)
```



RF Example – Interpretation (1 of 5)

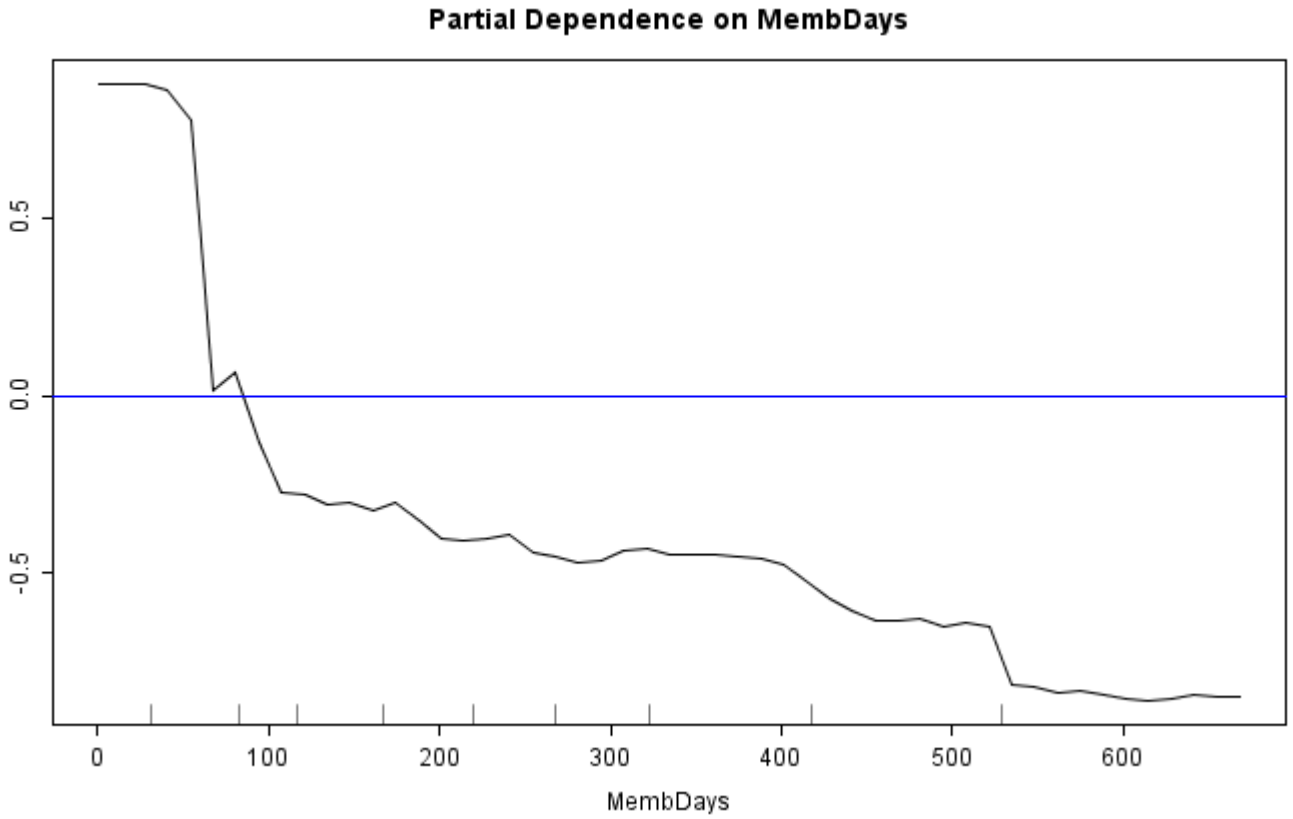
- Variable Importance Plot

```
> varImpPlot(Members.rf)
```



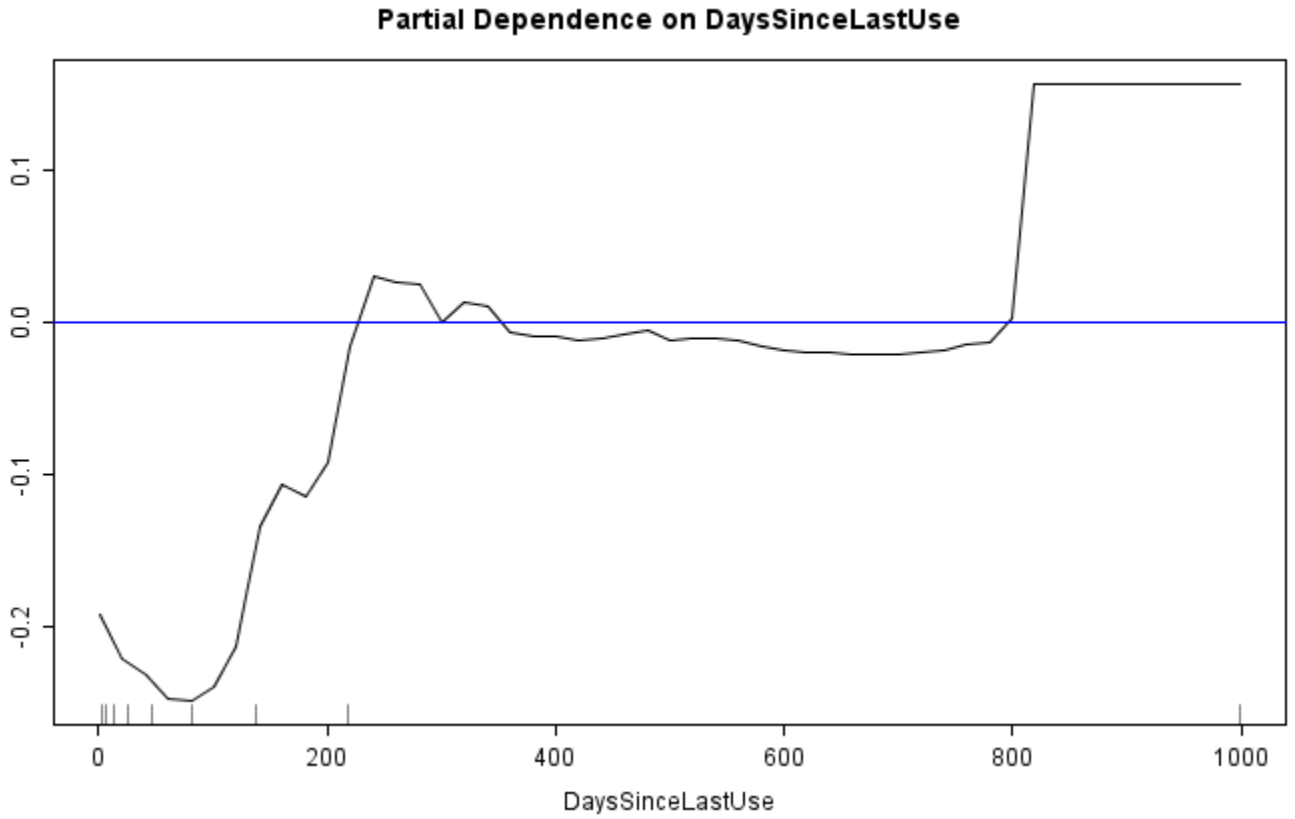
RF Example – Interpretation (2 of 5)

- RF Diagnostics – Partial Dependence 1
  - `partialPlot(Members.rf, Members[-1], MembDays)`
  - `abline(h=0, col = "blue")`



RF Example – Interpretation (3 of 5)

- RF Diagnostics – Partial Dependence 2
  - `partialPlot(Members.rf, Members[-1], DaysSinceLastUse)`
  - `abline(h=0, col = "blue")`



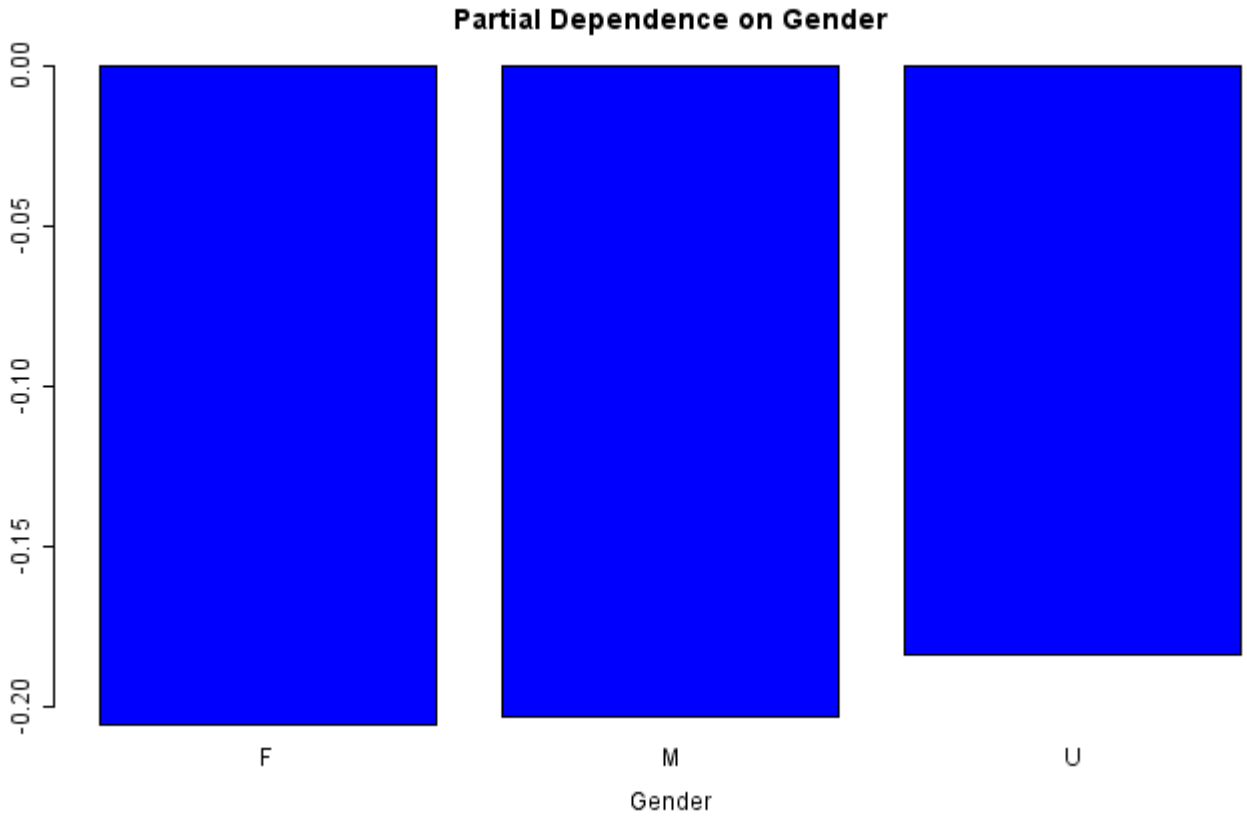
RF Example – Interpretation (4 of 5)

- RF Diagnostics – Partial Dependence 3
  - `partialPlot(Members.rf, Members[-1], Age)`

RF Example – Interpretation (5 of 5)

- RF Diagnostics – Partial Dependence 3

```
> partialPlot(Members.rf, Members[-1], Gender)
```



Loyalty Matrix

RF Example – Prediction (1 of 2)

- Need to do same variable selection & conditioning as we did for training set:

```
> MembersTest <- read.delim("Data/MemberTestSet.txt", row.names = "MembID")
> MembersTest <- subset(MembersTest, select = -c(FirstCkInDay,
LastCkInDay))
> MembersTest$DaysSinceLastUse[is.na(MembersTest$DaysSinceLastUse)] <- 999
> MembersTest$DaysSinceLastExtra[is.na(MembersTest$DaysSinceLastExtra)] <-
999
> MembersTest <- rfImpute(Status ~ ., data = MembersTest)
```

- Then we can use the "predict" method of our forest on the test data:

```
> MembersTest.pred <- predict(Members.rf, MembersTest[-1])
> str(MembersTest.pred)
```

```
 Factor w/ 2 levels "C","M": 2 2 2 2 2 1 1 2 2 1 ...
```

RF Example – Prediction (1 of 2)

- Some basic R gives the actual error:

```
> ct <- table(MembersTest[[1]], MembersTest.pred)
> cbind(ct, class.error = c(ct[1,2]/sum(ct[1,]), ct[2,1]/sum(ct[2,])))
```

```
  C   M class.error
C 511 295   0.3660050
M 144 951   0.1315068
```

```
> (ct[1, 2] + ct[2, 1]) / length(MembersTest$Status)  ## Test Set Error
```

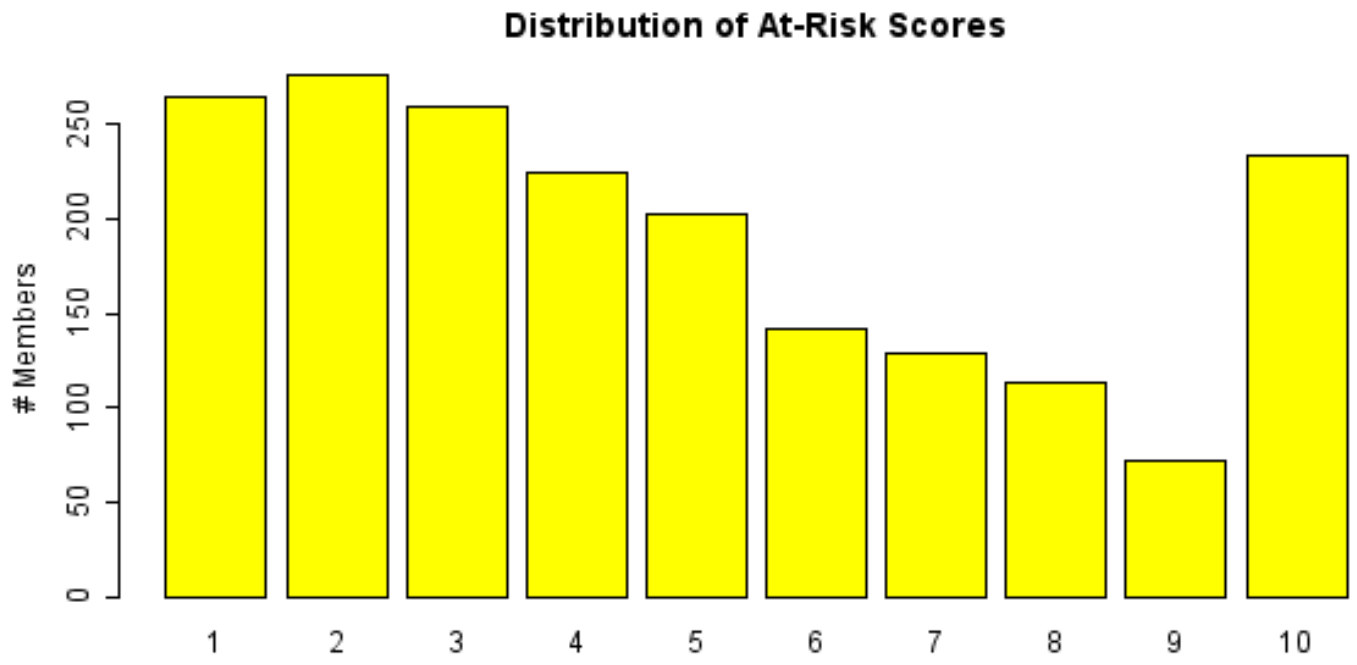
```
[1] 0.2309311
```

- Recall our original OOB error estimates:
  - 21% overall error rate.
  - 33% false positive
  - 13% false negative

RF Example – Scoring

- Need a score? Count the trees.

```
AtRiskScore <- floor(9.99999 * Members.rf$votes[, 1]) + 1
barplot(table(AtRiskScore), col = "yellow",
        ylab = "# Members", main = "Distribution of At-Risk Scores")
```



More Random Forests

- More capability in randomForest package
  - Regression Forest
  - Unsupervised Classification
  - Outlier measures
  - Prototypes
- Other Random Forests in R world
  - cforest in party package
    - Hothorn, Hornik & Zeileis; Vienna
  - varSelRF uses RF for variable selection
    - Ramón Díaz-Uriarte; Madrid

Conclusion - Random Forest Summary

- Has yielded practical results in number of cases
- Minimal tuning, no pruning required
- Black box, with interpretation
- Scoring fast & portable

Questions? Comments?

- Email JPorzak@LoyaltyMatrix.com
- Archive http://porzak.com/JimArchive/
- Call 415-296-1141
- Visit http://www.LoyaltyMatrix.com
- Come by at:
  580 Market Street, 6th Floor
  San Francisco, CA 94104

APPENDIX

R Setup for Tutorial

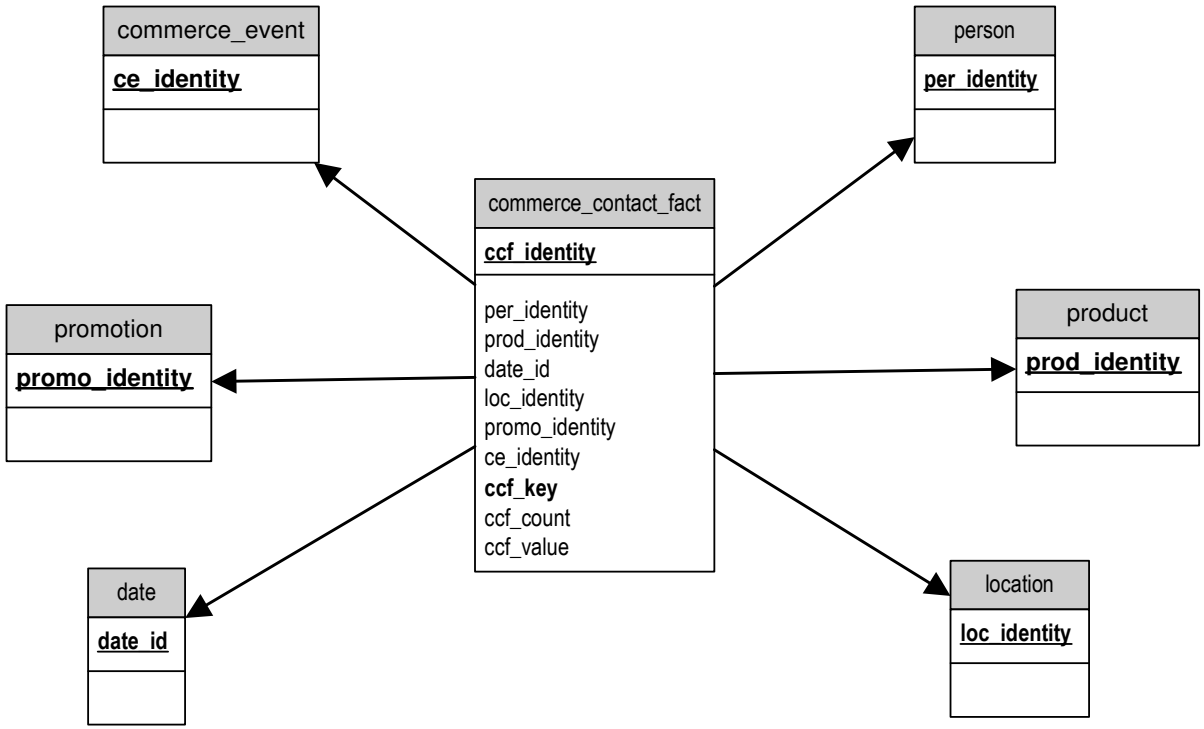
This is the setup I will be using during the tutorial, you may, of course, change OS, editor, paths to match your own preferences.

- Windows XP SP2 on 3GHz P4 w/ 1G RAM.
- R Version 2.4.0
- RWinEdt & WinEdt V5.4 or JGR
- Following packages will be used
  - randomForest
- Directory Structure
  - R's working directory & source code: C:\Projects\CIwR\R
  - Tutorial data loaded in: C:\Projects\CIwR\R\Data
  - Plots will be stored in: C:\Projects\CIwR\R\Plots
- Other tools I like to use
  - TextPad: www.TextPad.com
  - DbVisualizer: http://www.dbvis.com/products/dbvis/

- Download data/code from my archive: http://porzak.com/JimArchive/

Staging Data for Analysis – Star Schema

- RDBMS Datamart using a Star Schema
  - See Ralph Kimball: http://www.kimballgroup.com
  - Holds “Analysis Ready” data



Loyalty Matrix

Staging Data for Analysis – Moving to R

- Use RODBC to load directly from datamart

```
require(RODBC)
cODBC <- odbcConnect("KeyCustomers")                          # in Windows: odbcConnect("") works
myQuery <- readChar("../SQL/MyQuery.sql", nchars = 99999)     # use cat(myQuery) to view
MyDataFrame <- sqlQuery(cODBC, myQuery)
# Fix up datatypes, factors if necessary
MyDataFrame$DatePch <- as.Date(MyDataFrame$DatePch)
str(MyDataFrame)
head(MyDataFrame)
```

- Use SQL export & read.table
  - We'll use read.delim for tutorial (I like tab delimited)

```
KeyCustomers <- read.delim("Data/KeyCustomers.txt", row.names = "ActNum")
```

- Sampling large data sets
  - RANDOM table trick (two columns: integer identity & runif [0, 9999])

```
SELECT SUBT_ID, etc…
FROM NewSubscribers ns
  JOIN Random r
    ON r.identity_key = ns.SUBT_ID
   AND r.random <= 100  -- for 10% sample
```

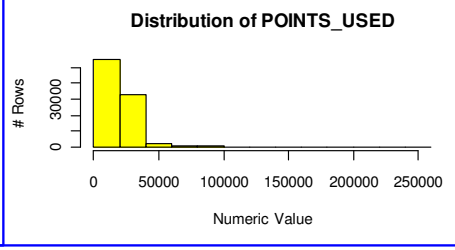
Profiling Raw Data in R

- Profile staged raw data to check assumptions about data made when defining problem

TriRaw . raw_redemptions . POINTS_USED 5 varchar(6)

	Rows	Nulls	Distinct	Empty	Numeric	Date
#	90,651	0	2,632	0	90,651	29,983
%	100.00	0.00	2.90	0.00	100.00	33.08

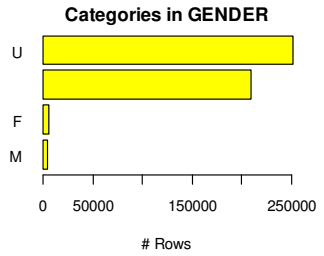
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
80	3,500	20,000	18,100	25,400	257,000



Head: 10000|10000|10000|10000|10000|10000

AMA_Stage . CUSTOMER . GENDER 14 varchar(8000)

	Rows	Nulls	Distinct	Empty	Numeric	Date
#	471,400	209,362	3	0	0	0
%	100.00	44.41	0.001	0.00	0.00	0.00



Head: NA|NA|NA|NA|NA|NA

Details in useR! 2006 talk
http://porzak.com/JimArchive/JimPorzak_RDataProfiling_useR2006_talk.pdf

Loyalty Matrix