Using a Random Forest model to predict enrollment

Ward Headstrom
Institutional Research
Humboldt State University
CAIR 2013

Overview

- Forecasting enrollment to assist University planning
- The R language
- The Random Forest model
- Binary Logistic Regression model
- Cautions and Conclusions

- The example I am going to use is projecting ***New*** enrollment. These techniques can easily be applied to predicting…
  - Retention
  - Graduation
  - Other future events

Simple enrollment projections

1) how many student enrolled last year?
2) enhance by breaking it down into subgroups
3) possibly use linear regressions (trends)

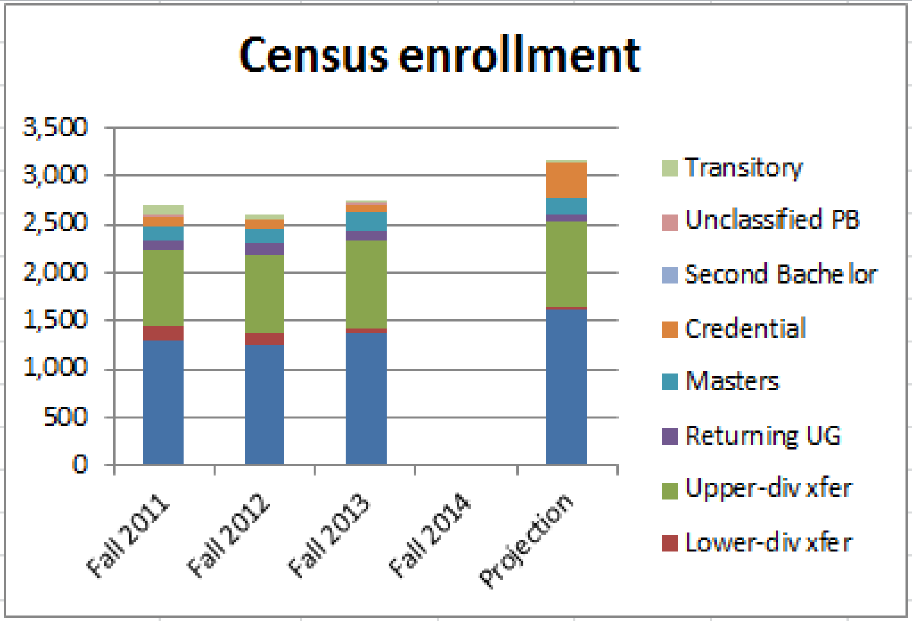
Applicant Counts Report Options					
Time frame	To-date	Final			
Semester	Fall	Spring	Summer	Academic year	
Cohort	Applicants	Admits	Active Admits	Confirmed	Registered

Final Fall Registered
report generated: 03-OCT-13

Applicant type	Fall 2007	Fall 2008	Fall 2009	Fall 2010	Fall 2011	Fall 2012	Fall 2013	Fall 2014
First-time UG	1,058	1,202	1,382	1,316	1,292	1,242	1,369	
Lower-div xfer	302	156	198	103	149	141	50	
Upper-div xfer	636	612	572	786	795	807	921	
Returning UG	110	102	99	84	108	108	96	
Masters	199	185	174	175	133	147	182	
Credential	168	117	124	118	111	107	92	
Second Bachelor	63	47	35	3	1	2	4	
Unclassified PB	17	7	10	4	5	2	5	
Transitory	176	183	30	71	99	43	30	
Totals	**2,729**	**2,611**	**2,624**	**2,660**	**2,693**	**2,599**	**2,749**	

4) enhance further by looking at to-date information

To-date Fall Applicants				
report generated: 05-NOV-13				
Applicant type	Fall 2011	Fall 2012	Fall 2013	Fall 2014
First-time UG	2,947	3,304	3,692	4,361
Lower-div xfer	1	132	13	10
Upper-div xfer	973	862	1,064	1,017
Returning UG	44	19	33	22
Masters	34	17	32	31
Credential	6	5	2	8
Second	10			
Unclassified PB	4		1	
Transitory				
Totals	4,019	4,339	4,837	5,449



Final Fall Registered								"to-date" yield rates		
report generated: 03-OCT-13					Fall 2014					
Applicant type	Fall 2011	Fall 2012	Fall 2013	Fall 2014	Projection	use this		Fall 2011	Fall 2012	Fall 2013
First-time UG	1,292	1,242	1,369		1,617	1,617		44%	38%	37%
Lower-div xfer	149	141	50		38	50		14900%	107%	385%
Upper-div xfer	795	807	921		880	880		82%	94%	87%
Returning UG	108	108	96		64	96		245%	568%	291%
Masters	133	147	182		176	182		391%	865%	569%
Credential	111	107	92		368	92		1850%	2140%	4600%
Second	1	2	4		4	4		10%	-	-
Unclassified PB	5	2	5		-	5		125%	-	500%
Transitory	99	43	30		34	30		-	-	-
Totals	2,693	2,599	2,749		3,182	2,956		67%	60%	57%

2014 projection = 2013 “to-date” yield * 2014 apps = 1,369/3,692*4,361 = 1,617

2014-15 Projection on 05-NOV-13

	Total Headcount			Resident FTE		
Student type	**Fall 2013**	**Fall 2014**	**%change**	**Fall 2013**	**Fall 2014**	**%change**
Continuing Undergrad	5,299	5,624	5.8%	4,874	5,173	5.8%
Returning Undergrad	96	96	0.0%	75	75	0.0%
First-time Undergrad	1,368	1,617	15.4%	1,277	1,509	15.4%
Transfer Undergrad	971	930	-4.4%	875	838	-4.4%
Continuing/Returning Postbac	249	256	2.6%	172	177	2.6%
New Postbac	278	283	1.8%	290	295	1.8%
Transitory	32	32	0.0%	20	20	0.0%
Totals	8,293	8,838	6.2%	7,583	8,088	6.2%

	Total Headcount			Resident FTE		
Student type	**Spring 14**	**Spring 15**	**%change**	**Spring 14**	**Spring 15**	**%change**
Continuing Undergrad	6,976	7,457	6.5%	6,348	6,786	6.5%
Returning Undergrad	21	21	0.0%	19	19	0.0%
First-time Undergrad	39	39	0.0%	35	35	0.0%
Transfer Undergrad	408	408	0.0%	371	371	0.0%
Continuing/Returning Postbac	430	439	2.1%	391	400	2.1%
New Postbac	29	29	0.0%	26	26	0.0%
Transitory	39	39	0.0%	35	35	0.0%
Totals	7,941	8,432	5.8%	7,226	7,673	5.8%

But what about…

- Why applicant yield might not be the best predictor:
  - Admits more likely to enroll
  - Confirms more likely to enroll
  - Denied or withdrawn will not enroll
  - Housing deposits may be good indicator of intent
  - Local applicants more likely than distant applicants
  - Certain majors or ethnicities may be more likely to enroll
  - Do this year’s applicants look like last year’s?
- Ideally, we would like to use all the data we have about applicants to predict how likely they are to enroll.
  - Variables: demographics, academics, actions to-date
  - Model 1: Random Forest
  - Model 2: Binary logistic regression

The language R

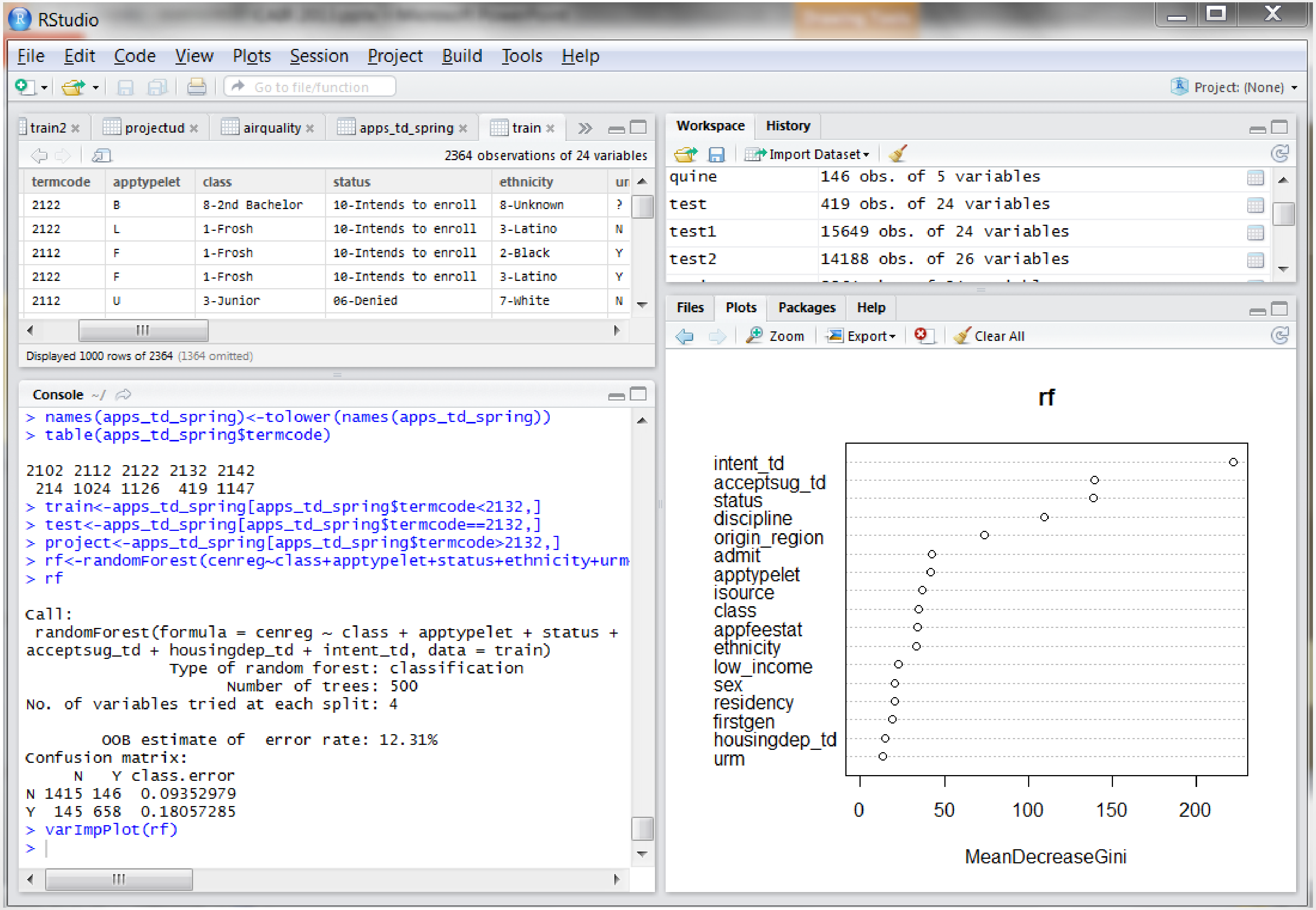
CAIR comment: an emphasis on R would be “limiting to institutions that used other software”.

- The first (only?) implementation of Random Forest models
- R is open source – free to use
  - http://cran.us.r-project.org/
  - http://www.rstudio.com/ide/download/desktop
- Many online tutorials:
  - http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf
  - http://bioinformatics.knowledgeblog.org/2011/06/21/using-r-a-guide-for-complete-beginners/
  - https://www.coursera.org/course/compdata
- www.researchgate.net/post/Which_is_better_R_or_SPSS

R and RStudio overview

- Function-based: function(data,options)
- Case-specific language
- 4 panes – help, history, import dataset, ***packages***
- Object types: data.frame, vector, scalars, factor, models
- Useful commands:
  - command line console can be used as calculator
  - assignment -> or <-
  - functions: na.omit(), summary(), table(), tolower()
  - subsets: dataframe[row select, column select]
  - graphics: hist(), plot()
  - library() , especially library(randomForest)

RStudio


RStudio
File Edit Code View Plots Session Project Build Tools Help
Go to file/function
Project: (None)
train2 projectud airquality apps_td_spring train
2364 observations of 24 variables
termcode apptypelet class status ethnicity ur
2122 B 8-2nd Bachelor 10-Intends to enroll 8-Unknown ?
2122 L 1-Frosh 10-Intends to enroll 3-Latino N
2112 F 1-Frosh 10-Intends to enroll 2-Black Y
2122 F 1-Frosh 10-Intends to enroll 3-Latino Y
2112 U 3-Junior 06-Denied 7-White N
Displayed 1000 rows of 2364 (1364 omitted)
Workspace History
Import Dataset
quine 146 obs. of 5 variables
test 419 obs. of 24 variables
test1 15649 obs. of 24 variables
test2 14188 obs. of 26 variables
Files Plots Packages Help
Zoom Export Clear All
Console ~/
> names(apps_td_spring)<-tolower(names(apps_td_spring))
> table(apps_td_spring$termcode)
2102 2112 2122 2132 2142
214 1024 1126 419 1147
> train<-apps_td_spring[apps_td_spring$termcode<2132,]
> test<-apps_td_spring[apps_td_spring$termcode==2132,]
> project<-apps_td_spring[apps_td_spring$termcode>2132,]
> rf<-randomForest(cenreg~class+apptypelet+status+ethnicity+urm
> rf
Call:
randomForest(formula = cenreg ~ class + apptypelet + status +
acceptsug_td + housingdep_td + intent_td, data = train)
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 4
OOB estimate of error rate: 12.31%
Confusion matrix:
N Y class.error
N 1415 146 0.09352979
Y 145 658 0.18057285
> varImpPlot(rf)
>
rf
intent_td
acceptsug_td
status
discipline
origin_region
admit
apptypelet
isource
class
appfeestat
ethnicity
low_income
sex
residency
firstgen
housingdep_td
urm
0
50
100
150
200
MeanDecreaseGini

Data files

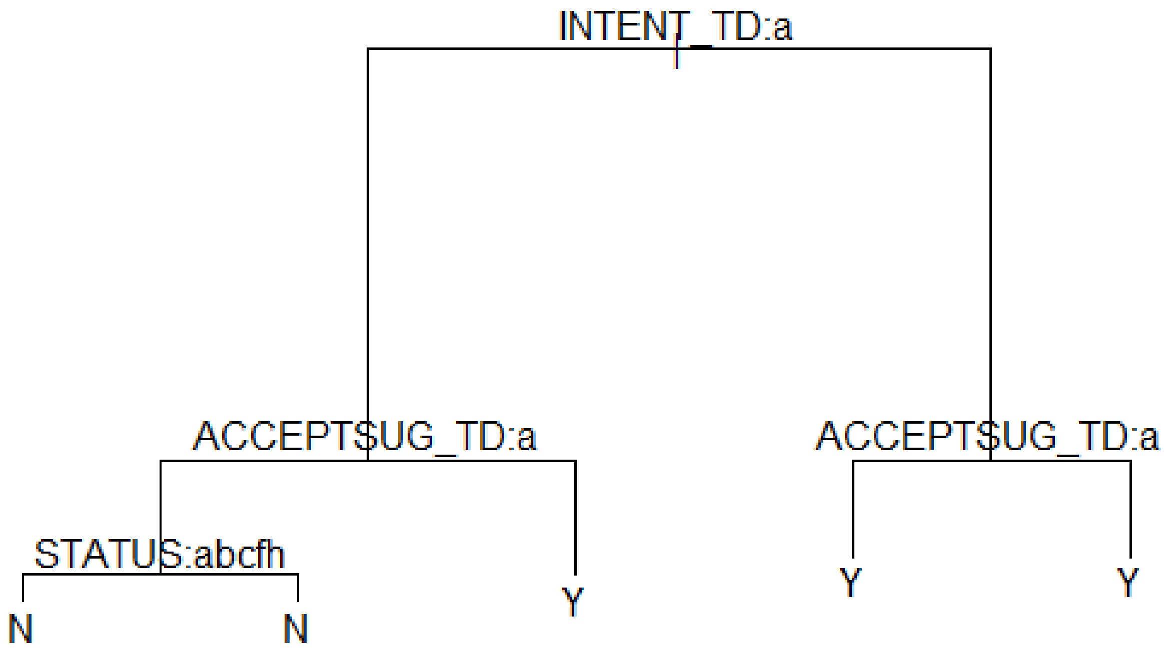
- All the data fields you think might help predict yields
  - Major discipline
  - Region of origin
  - Sex
  - Ethnicity
  - Academic preparation
- Actions “to-date”
  - Accepted SUG
  - Confirmed intent to enroll
  - Paid housing deposit
- Institutional actions
  - Admit
  - Deny/cancel

Import data into R

```
> apps_td_spring <- read.csv("C:/Users/ward/Google Drive/IRP/CAIR 2013/apps_td_spring.csv")
> summary(apps_td_spring)
       ID                 TERMCODE      APPTYPELET            CLASS
 Min.    : 10002145   Min.    :2102   U        :2515   3-Junior      :2424
 1st Qu.: 11272781   1st Qu.:2112   R        : 426   1-Frosh       : 520
 Median : 12152872   Median :2122   F        : 339   4-Senior      : 463
 Mean    :423225308   Mean    :2125   L        : 311   2-Sophomore   : 221
 3rd Qu.:946602927   3rd Qu.:2142   M        : 242   6-Masters     : 156
 Max.    :953487425   Max.    :2142   N        :  44   E-Extended Ed:  86
                                      (Other):  53   (Other)       :  60
                  STATUS              ETHNICITY     URM        SEX           ORIGIN_REGION
 03-Applied          :1168   7-White       :2036   ?: 329   F:1957   7-Los Angeles:926
 10-Intends to enroll:1047   3-Latino      : 619   N:2383   M:1854   1-Local      :801
 07-Admitted         : 749   6-Two or more: 330   Y:1218   U: 119   2-Northern CA:412
 06-Denied           : 286   8-Unknown     : 309                     3-SF Bay     :409
 05-Complete         : 258   4-Asian       : 250                     X-Other state:289
 05-In Review        : 196   2-Black       : 243                     6-Central CA :242
 (Other)             : 226   (Other)       : 143                     (Other)      :851
   RESIDENCY      EXCADMIT CENREG   LOW_INCOME FIRSTGEN       HS_GPA       INTENT_TD ADMIT
 R       :3403    :  13   N:2937   N:2901       : 216   Min.    :1.340   N:2752    N:2062
 N       : 168   N:3893   Y: 993   Y:1029      N:1705   1st Qu.:2.680   Y:1178    Y:1868
 X       : 147   Y:  24                        Y:2009   Median :3.000
 F       :  93                                          Mean    :3.001
 WUE     :  87                                          3rd Qu.:3.350
         :  16                                          Max.    :4.330
 (Other):  16                                           NA's    :3609
 ACTIVE    EOP_INTEREST HOUSINGDEP_TD            ISOURCE          APPFEESTAT
 N: 498     :   2       N:3815        APPLICATION        :2579   CC      :2079
 Y:3432    N:2645       Y: 115        STUDENT INITIATED:1011   FW      :1447
           Y:1283                     TRAVEL             : 117   PD      : 142
                                      DIRECT MAIL        :  84   ?       : 116
                                      CAMPUS VISITOR     :  76   NP      : 113
                                      REFERRAL           :  53           :  13
                                      (Other)            :  10   (Other):  20
```

Decision Trees

```
> tr<-tree(CENREG~ACCEPTSUG_TD+INTENT_TD+STATUS,data=apps_td_spring)
> plot(tr)
> text(tr)
```



Random Forest Model

- Developed by Leo Brieman and Adele Cutler
- Plan: grow a random forest of 500 decision trees
  - randomForest(cenreg~variable1+variable2+…,data=train)
  - Randomly picks fields for each tree
  - Randomly selects rows to exclude from each tree
- Measure of variable importance
- Out Of Box estimate of error rate and Confusion matrix

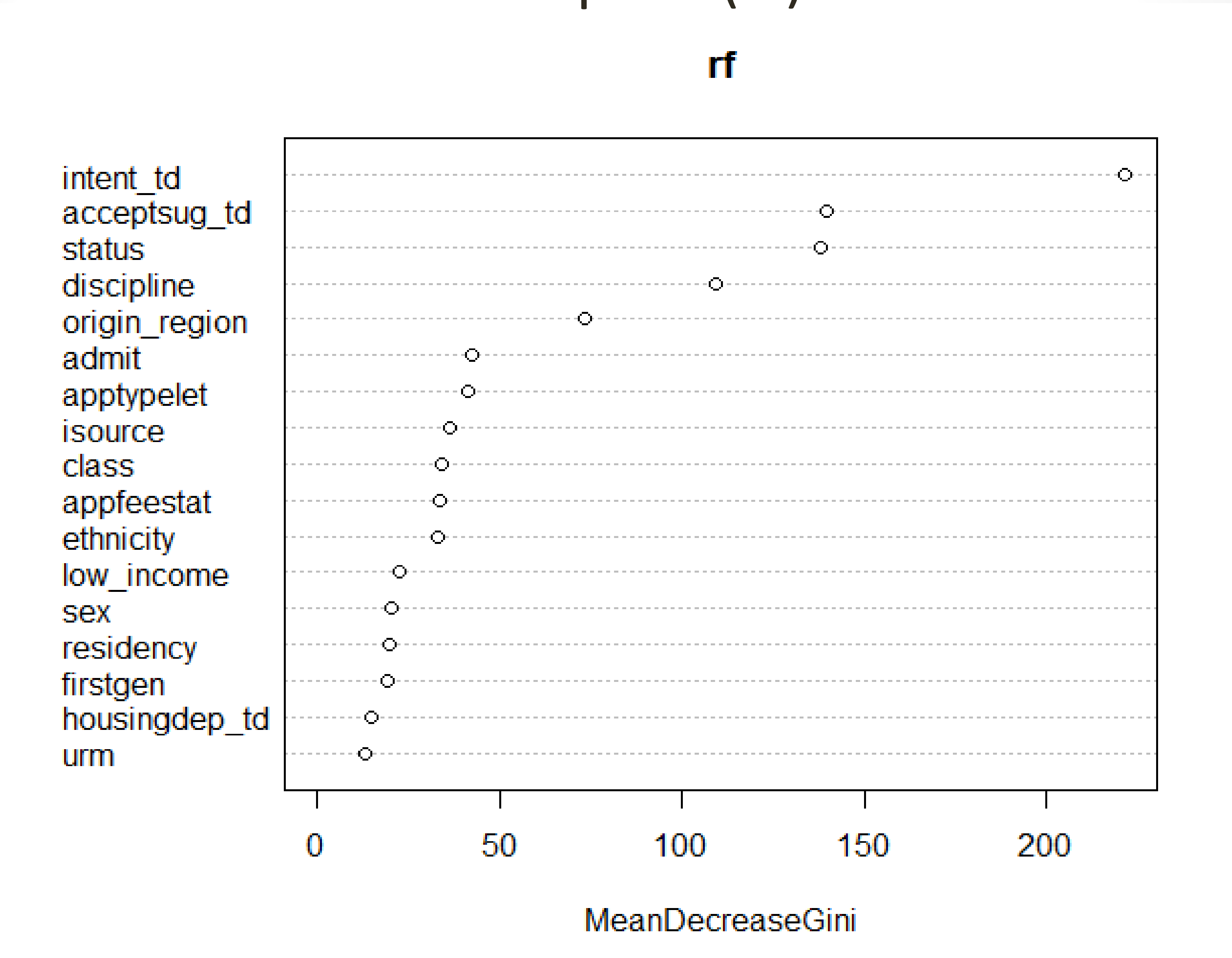
```
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 12.52%
Confusion matrix:
     N   Y class.error
N 1414 147   0.0941704
Y  149 654   0.1855542
```

- Run new data through all 500 trees and let them vote

Random Forest model of applicant yield

```
> names(apps_td_spring)<-tolower(names(apps_td_spring))
> table(apps_td_spring$termcode)

2102 2112 2122 2132 2142
 214 1024 1126  419 1147
> train<-apps_td_spring[apps_td_spring$termcode<2132,]
> test<-apps_td_spring[apps_td_spring$termcode==2132,]
> project<-apps_td_spring[apps_td_spring$termcode>2132,]
> rf<-randomForest(cenreg~class+apptypelet+status+ethnicity+urm
> rf

Call:
 randomForest(formula = cenreg ~ class + apptypelet + status +
acceptsug_td + housingdep_td + intent_td, data = train)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 12.31%
Confusion matrix:
     N   Y class.error
N 1415 146  0.09352979
Y  145 658  0.18057285
> varImpPlot(rf)
```

varImpPlot(rf)



1st tree in Random Forest

```
> head(getTree(rf,1,labelVar=TRUE),7)
  left daughter right daughter    split var split point status prediction
1             2              3        admit           1      1       <NA>
2             4              5   apptypelet         210      1       <NA>
3             6              7        class          21      1       <NA>
4             8              9   discipline     3543222      1       <NA>
5            10             11 acceptsug_td           1      1       <NA>
6            12             13       status         192      1       <NA>
7            14             15       status         320      1       <NA>
> tail(getTree(rf,1,labelVar=TRUE),7)
    left daughter right daughter  split var split point status prediction
545           550            551 discipline     4177919      1       <NA>
546             0              0       <NA>           0     -1          N
547             0              0       <NA>           0     -1          N
548             0              0       <NA>           0     -1          Y
549             0              0       <NA>           0     -1          Y
550             0              0       <NA>           0     -1          Y
551             0              0       <NA>           0     -1          N
```

For categorical predictors, the splitting point is represented by an integer, whose binary expansion gives the identities of the categories that goes to left or right. For example, if a predictor has four categories, and the split point is 13. The binary expansion of 13 is (1, 0, 1, 1) (because *13 = 1*2^0 + 0*2^1 + 1*2^2 + 1*2^3*), so cases with categories 1, 3, or 4 in this predictor get sent to the left, and the rest to the right.

Testing and making a Projection

```
> test$rf<-predict(rf,test)
> table(test$cenreg,test$rf)

      N   Y
  N 196  33
  Y  21 169
```

```
> table(test$cenreg)

  N   Y
229 190
> 190/(229+190)
[1] 0.4534606
```

```
> table(project$apptypelet,predict(rf,project))

      N   Y
  B   2   1
  C   0   0
  F  81  36
  L  51   3
  M  75   9
  N   2   5
  P   0   0
  R  90  43
  U 506 243
```

```
> table(predict(rf,project))

  N   Y
807 340
```

Random Forest projects that 42% of current Spring apps will enroll, compared to 45% of last year's apps to-date and 34% of training years'.

Binary Logistic Regression

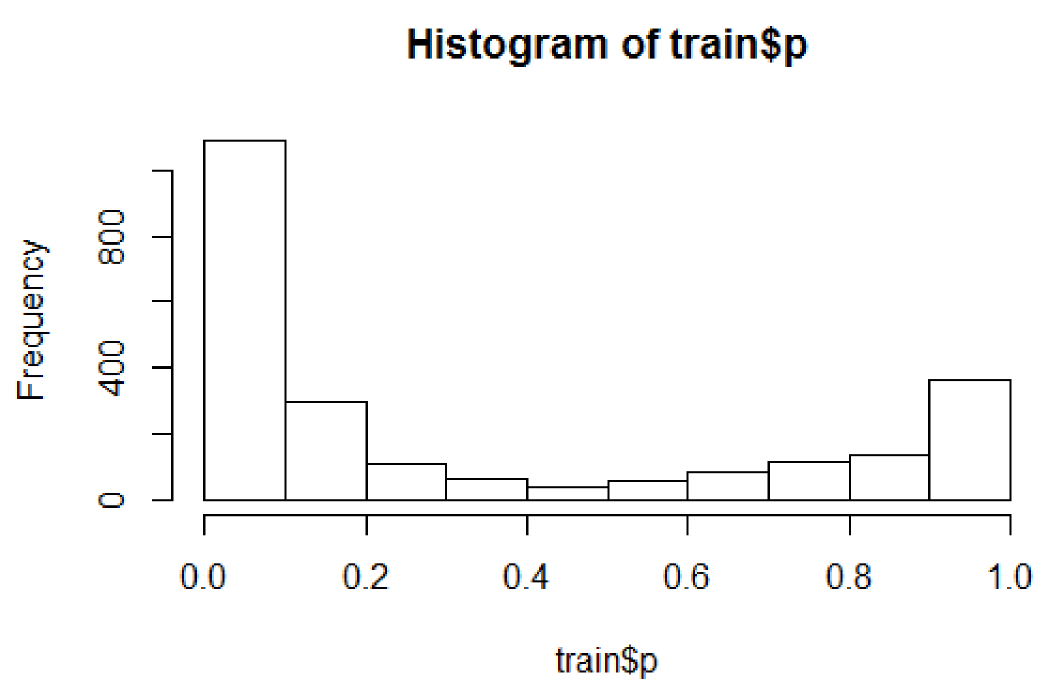
- p(x) is the probability that x will occur, where x is a binary object (Y/N, 1/0, true/false)
- $\log\left(\frac{p(x)}{1-p(x)}\right) = B_0 + B_1 * X_1 + B_2 * X_2 + B_3 * X_3 + \cdots$
- B_n represents calculated coefficients
- X_n represents the value of dependent variables
- Break up factor variables into many terms where X_n is 1 or 0
- Can manipulate the result to return the probability (between 0 and 1) that x will occur, given the state of a particular set of dependent variables.
- Difficult to predict outcome of a single individual
- Can sum probabilities to estimate total

Binary logistic regression model of applicant yield

```
> blr<-glm(cenreg~status+acceptsug_td+intent_td,data=train,family=binomial)
> summary(blr)

Call:
glm(formula = cenreg ~ status + acceptsug_td + intent_td, family = binomial,
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8129  -0.5208  -0.4253   0.1125   2.6771

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -2.3575     0.1225 -19.246  < 2e-16 ***
status04-Withdrew app       -0.4305     0.7284  -0.591   0.5545
status05-Complete            1.0736     0.2348   4.573 4.80e-06 ***
status05-In Review           0.4282     0.3183   1.345   0.1785
status06-Denied             -1.1979     0.5217  -2.296   0.0217 *
status07-Admitted            0.8117     0.1709   4.748 2.05e-06 ***
status08-Not coming        -14.2086   438.0949  -0.032   0.9741
status10-Housing deposit     0.5734     0.8411   0.682   0.4955
status10-Intends to enroll   1.1399     0.2370   4.809 1.52e-06 ***
acceptsug_tdY                6.2944     1.0096   6.234 4.54e-10 ***
intent_tdY                   2.3712     0.1984  11.954  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3029.8  on 2363  degrees of freedom
Residual deviance: 1643.7  on 2353  degrees of freedom
AIC: 1665.7

Number of Fisher Scoring iterations: 15
```

```
> anova(blr,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: cenreg

Terms added sequentially (first to last)


              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           2363     3029.8
class          8    80.07      2355     2949.7 4.743e-14 ***
apptypelet     5    30.78      2350     2919.0 1.035e-05 ***
status         8   822.38      2342     2096.6 < 2.2e-16 ***
ethnicity      7    29.78      2335     2066.8  0.000104 ***
urm            2     5.08      2333     2061.7  0.078694 .
sex            2     2.29      2331     2059.4  0.317455
origin_region 11    38.74      2320     2020.7 5.863e-05 ***
residency      6    43.73      2314     1977.0 8.357e-08 ***
low_income     1    42.23      2313     1934.7 8.131e-11 ***
firstgen       2    12.23      2311     1922.5  0.002211 **
admit          0     0.00      2311     1922.5
isource        8    23.25      2303     1899.2  0.003060 **
appfeestat     9    27.71      2294     1871.5  0.001066 **
discipline    21    37.19      2273     1834.3  0.015998 *
acceptsug_td   1   303.95      2272     1530.4 < 2.2e-16 ***
housingdep_td  1    65.38      2271     1465.0 6.176e-16 ***
intent_td      1   157.64      2270     1307.4 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> train$p<-predict(blr,train,type="response")

> table(round(train$p,1),train$cenreg)

        N   Y
  0   729  13
  0.1 501  43
  0.2 128  45
  0.3  55  19
  0.4  25  24
  0.5  18  29
  0.6  26  44
  0.7  32  71
  0.8  27 112
  0.9  19  79
  1     1 324
>
> hist(train$p)
```



BLR model – testing and projecting

```
>
> sum(test$cenreg=="Y")
[1] 190
> test$p<-predict(blr,test,type="response")
> sum(test$p)
[1] 182.7157
```

```
> project$p<-predict(blr,project,type="response")
> sum(project$p)
[1] 324.4795
```

Binary Logistic Regression predicted 324 of current Spring applicants will enroll, compared to 340 projected by Random Forest model.

Cautions and Conclusions

- Null or new values in variables will cause problems
- Beware of to-date variables (e.g. intent_td). Make sure that procedures have not changed in a way that will affect behavior.
- R is a very powerful tool which can be very useful if you are willing to invest some time learning it.
- Multivariate models ***may*** improve the accuracy of your predictions. Corroborate with simple models and consultation with involved staff.

Questions? Comments?

This presentation:

www.humboldt.edu\irp\presentations.html

My email:

Ward.Headstrom@humboldt.edu