Modeling Attrition
A comparison of techniques

Bill Romine - LLNL
John Pantano - LANL
June, 2013
Privacy & Legal Notice

Disclaimer
This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

Auspices Statement
This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.
Presentation Outline

1) Attrition prediction problem
2) Institutional context
3) Historical precedents
   a) Actuarial practice
   b) Los Alamos’ ongoing effort
4) Data mining overview
5) Comparison study
6) References and software
7) Wrap-up conclusions
Workforce attrition prediction

• An organization’s workforce data can be sampled by a series of roll call audits taken periodically at a fixed time interval. Each audit records individual x’s vital statistics along with a label according to employment status:
  – ACT: if x neither retired nor terminated during the interval and remains active
  – RET: if x stopped working and began collecting accrued retirement benefits
  – TRM: if x left employment with the organization w/ or w/out vested benefits

• To maintain a stable workforce the population decrements for RET and TRM must be offset by an equivalent level of replacement hiring.

• Attrition prediction provides a tool for workforce planning and management.
Institutions

• Los Alamos
  – Founded in 1943 as site Y of the Manhattan Project under the leadership of J. Robert Oppenheimer of the University of California
  – Currently about 10,000 employees

• Livermore
  – Founded in 1952 under the leadership of Ernest O. Lawrence of the University of California
    (Lawrence’s invention, the cyclotron, was used during the Manhattan Project to enrich uranium at the Oakridge site)
  – Currently about 7,000 employees

http://www.lanl.gov/history/

http://education.llnl.gov/archives/
Attrition in the context of career employment

• Historically* personnel management policies were determined by the University of California.

• Individuals were hired into one of 4 appointment types:
  – Student / Post-Doc (not more than 3 years)
  – Limited Term (not more than 7 years)
  – Career Indefinite (no expiration time)
  – Non-career (intended for post-retirement part-time returnees)

• Attrition prediction is restricted to Career employees who represent about 80% of the total workforce.
  – Retirement was under the UCRP defined benefit pension plan.
  – Non-retirement termination includes voluntary separation, involuntary separation, temporary/permanent disability and death.

* Prior to privatization in 2006 (Los Alamos) and 2007 (Livermore)
Actuarial practice\textsuperscript{[1]}: Age, Service, interactions

Models are derived by multinomial logistic regression from data representing 1.7 million life years of exposure during years 1994-2000 for 112 separate defined benefit pension plans.

\textsuperscript{[1]} Frees, Edward W., *Pension Plan Termination and Retirement Study*, Society of Actuaries, 2003

\* Pr( event during year | active at beginning of year )
Ongoing work at Los Alamos
Regression methods for predicting attrition

- Goal is to find the probability of an individual terminating

- Models can be based on:
  - Logistic regression
  - Ordinary Least Squares (OLS) regression

- Common data treatment
  - dichotomous dependent variable
  - categorical predictors entered as dummy variables
Attrition outcome variable

- Dependent variable (Y) is a dichotomous variable:
  - 0 = has not terminated, and
  - 1 = has terminated. If there is a large enough sample, this can be treated as ratio data for the purposes of OLS regression.
Predictor variables in linear model

- **Dynamic variables (all continuous)**
  - Age
  - Service
  - \(P(Age, Svc), P\) polynomial at most degree 3 in Age & Service
  - moSalary

- **Static variables (all categorical)**
  - Gender
  - Ethnicity
  - Job Classification: R&D, Manager, Professional, Tec, Support
  - Type of Organization: STE, Weapons, Operations, Business
  - Local birth: Born in New Mexico
  - Transition annuity status: TCP2 (401K with match), TCP1 (Defined Benefits), none
  - Term Pattern (based on economic and internal conditions possibly affecting terminations)
OLS vs. logistic regression compared

Equations in their simplest form

- Ordinary least Squares (OLS)
  \[ Y = a + \sum b_i x_i \]

- Logistic Regression
  - Focus on the Probability (P) of Y = 1
    - \( \text{odds} = P / (1-P) \)
  - In logistic regression the dependent variable is a logit (natural log of the odds) function.
    - \( \ln(\text{odds}) = a + \sum b_i x_i \) or \( \ln(P/(1-P)) = a + \sum b_i x_i \)
    - which leads to
      \[ P = \frac{e^{(a+\sum b_i x_i)}}{1+e^{(a+\sum b_i x_i)}} \]

Both techniques will predict probability but the logistic model:

- Is designed to predict a dichotomous outcome.
- Is stated in terms of the probability of the outcome occurring and will keep the probability between 0 and 1
OLS, ordinary least squares, in practice

- Although logistic regression theoretically fits the attrition model scenario better, ordinary least squares has proven to be a good technique in practice.
- With large sample sizes results are similar to logistic regression.
- Treats dichotomous variables as continuous which works for large sample sizes.
- Experience at LANL is that OLS regression has worked well for predicting terminations in the short term (1 yr) where age and years of service are important factors.
Steps in Attrition Modeling

- Logistic regression to assist in identifying significant factors
  - Retirements: Age, years of service, retirement plan, organization type, term pattern, born in NM, gender and job classification.
  - Non-Retirement attrition: Age, years of service, retirement plan, organization type, term pattern, born in NM, ethnicity and job classification.

- Determine nature of relationships.

- Build and test models:
  - Logistic regression
  - Least squares regression model with categorical variables entered as dummy variables
The nature of the relationships indicate a polynomial rather than a linear relationship between age, years of service and termination rates.

P(Age,Svc) where P is a polynomial of degree at most 3 in Age & Service is added to the analysis.
Models Used for 2013 – 2017 Projections

- Ordinary Least Squares Regression for modeling terminations with one year.
- Logistic Regression for modeling terminations with the next five years.
Testing 1 year prediction equations historically

Significant events:
- 2004 – safety issues and stand down
- 2005 – contract negotiations
- 2006 – contract transition
- 2007 – first full year under new management
- 2008 – Self Selection Program
- 2009 – 2011- Economy
- 2012 – Voluntary Separation Program
Data mining overview

• Single model techniques
  – Stepwise regression
    • Best at following temporal patterns
    • Smooth model facilitates calculation of individual probabilities
  – Trees (both deviance, CART)
    • Using cross-validation & complexity pruning can be highly accurate when well tuned or when driven interactively

• Ensemble or black box predictors
  – Random forests of bagged C&R trees
    • Capable of regression level accuracy without tuning
  – Deterministic and stochastic gradient boosting
    • So far hasn’t improved prediction in experiments with JMP Pro
An academic example of classification tree
Iris flower: 3 erect petals & 3 horizontal sepals

setosa: Wild Flag Iris
occurs from northwest BC to Alaska

versicolor: Harlequin Blue Flag Iris
occurs from Minnesota to Virginia

virginica: Southern Blue Flag Iris
occurs from Virginia to eastern Texas
Data begins intermixed and ends in nearly pure labeled regions

Red = setosa
Blue = virginica
Green = versicolor
Individual tree algorithms

- Next split choice criteria
  - maximum reduction in deviance* over all allowed splits
    * Each node i and class k, \( n_{ik} \) drawn from \( p_{ik} \) multinomial probability distribution over classes deviance = \( -2 \sum_k n_{ik} \log p_{ik} \)
  - maximum reduction in average impurity = \( \sum_{j \neq k} p_{ij} p_{ik} \)

- Recursive partition stopping rules (if split doesn’t reduce by)
  - deviance threshold
  - complexity parameter threshold

- With noisy data (i.e. class distributions overlap) recursive partitioning adapts too well to features of learning sample

- Pruning provides an analogy to stepwise selection in regression
Individual tree fitting applied to Iris data

> print(colnames(iris))
> ir.tf<-tree(as.factor(Species)~.,data=iris,split="deviance")
> print(ir.tf,digits=1)
node), split, n, deviance, yval, (yprob)
 * denotes terminal node

1) root 150 300 setosa ( 0.33 0.33 0.33 )
   2) Petal.Length < 2.45 50 0 setosa ( 1.00 0.00 0.00 ) *
   3) Petal.Length > 2.45 100 100 versicolor ( 0.00 0.50 0.50 )
      6) Petal.Width < 1.75 54 30 versicolor ( 0.00 0.91 0.09 )
         12) Petal.Length < 4.95 48 10 versicolor ( 0.00 0.98 0.02 )
            24) Sepal.Length < 5.15 5 5 versicolor ( 0.00 0.80 0.20 ) *
            25) Sepal.Length > 5.15 43 0 versicolor ( 0.00 1.00 0.00 ) *
   13) Petal.Length > 4.95 6 8 virginica ( 0.00 0.33 0.67 ) *
    7) Petal.Width > 1.75 46 10 virginica ( 0.00 0.02 0.98 )
    14) Petal.Length < 4.95 6 5 virginica ( 0.00 0.17 0.83 ) *
    15) Petal.Length > 4.95 40 0 virginica ( 0.00 0.00 1.00 ) *

> print(summary(ir.tf))

Classification tree:
tree(formula = as.factor(Species) ~ ., data = iris, split = "deviance")
Variables actually used in tree construction:
Number of terminal nodes: 6
Residual mean deviance: 0.1253 = 18.05 / 144
Misclassification error rate: 0.02667 = 4 / 150
Classification of Iris species
based on petal length, width; sepal length
Random forest algorithm

- Draw a bootstrap sample from the data w/ replacement
  - About 2/3rd of data are “in the bag” used for training, and
  - The remaining 1/3rd are outside-of-bag and used for testing.

- Grow a classification tree by best splits on a small number m of randomly chosen predictor variables.
  - Best splits found that most increase node purity defined by the Gini index.
  - Trees are grown to the extent possible. There is no pruning.

- Estimate the error of prediction by comparing tree classification to actual class labeling for the OOB data.

- Iterate to build the forest.
  The only parameters are the number of trees in the forest and the number of predictor variables to be chosen to build the trees in the forest.
Random forest applied to Iris data

> set.seed(71)
> ir.rf <- randomForest(Species ~ ., data = iris, importance = TRUE, proximity = TRUE)

> print(ir.rf)

Call:
randomForest(formula = Species ~ ., data = iris, importance = TRUE, proximity = TRUE)
  Type of random forest: classification
  Number of trees: 500
No. of variables tried at each split: 2

OOB estimate of error rate: 4%

Confusion matrix:

<table>
<thead>
<tr>
<th></th>
<th>setosa</th>
<th>versicolor</th>
<th>virginica</th>
<th>class.error</th>
</tr>
</thead>
<tbody>
<tr>
<td>setosa</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>versicolor</td>
<td>0</td>
<td>47</td>
<td>3</td>
<td>0.06</td>
</tr>
<tr>
<td>virginica</td>
<td>0</td>
<td>3</td>
<td>47</td>
<td>0.06</td>
</tr>
</tbody>
</table>

> round(importance(ir.rf),2)

<table>
<thead>
<tr>
<th></th>
<th>setosa</th>
<th>versicolor</th>
<th>virginica</th>
<th>MeanDecreaseAccuracy</th>
<th>MeanDecreaseGini</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sepal.Length</td>
<td>1.49</td>
<td>1.53</td>
<td>1.93</td>
<td>1.32</td>
<td>9.89</td>
</tr>
<tr>
<td>Sepal.Width</td>
<td>1.11</td>
<td>0.10</td>
<td>1.27</td>
<td>0.71</td>
<td>2.27</td>
</tr>
<tr>
<td>Petal.Length</td>
<td>3.81</td>
<td>4.42</td>
<td>4.19</td>
<td>2.52</td>
<td>42.77</td>
</tr>
<tr>
<td>Petal.Width</td>
<td>3.76</td>
<td>4.35</td>
<td>4.28</td>
<td>2.50</td>
<td>44.33</td>
</tr>
</tbody>
</table>
Random forest: LLNL retirements FY2002

OOB estimate of error rate: 9.01%
Confusion matrix:

<table>
<thead>
<tr>
<th></th>
<th>ACT</th>
<th>RET</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>2119</td>
<td>51</td>
</tr>
<tr>
<td>RET</td>
<td>163</td>
<td>43</td>
</tr>
</tbody>
</table>

\[
\text{class.error} = \frac{51}{(2119 + 51)} \quad \text{and} \quad \frac{163}{(163 + 43)}
\]

- Notice that over 3/4 of the time training trees will misclassify RET. This reflects the balance between the relative sizes of ACT & RET. Breiman adjusts weights for the classes beginning with 1 / training size.

- The error rate for all the test classifications ~ 9%.

Random Forests reduce variance (increasing confidence in predictions) not bias.
Random forest generates an importance ranking

- Termination related variables like SVC, SVC^2, SVC^3 rank lower.
- Retirement related variables like AGE, AGE^2, AGE^3 rank higher.
- This ranking is consistent with the regression based models.

LLNL retirement data
Comparison study

Decision rule and partition tree

Parameter Estimates

| Term      | Estimate  | Std Error | t Ratio | Prob>|t| |
|-----------|-----------|-----------|---------|-----------------|
| Intercept | -0.051820 | 0.013182  | -3.92   | <.0001*         |
| AGE       | 0.000794  | 0.000232  | 3.21    | 0.0043*         |
| SVC       | 0.0474357 | 0.0035344 | 13.42   | <.0001*         |
| moPay     | 3.4123e-6 | 6.812e-7  | 4.94    | <.0001*         |
| AGE*SVC   | -0.0021   | 0.000131  | -15.99  | <.0001*         |
| SVC*AGEsq | 2.2847e-5 | 1.215e-6  | 18.80   | <.0001*         |
| PHD       | -0.015654 | 0.002521  | -6.16   | <.0001*         |
| MD        | 0.0568889 | 0.028945  | 2.11    | 0.0348*         |

Status =

-0.051823183954
+0.00079356365926*AGE
+0.04743573501897*SVC
+0.00000341226716*moPay
+0.0002100256355*AGE*SVC
+0.000002264705291*SVC*AGEsq
+0.0156409386963*PhD
+0.056888323807394*MD
Regression approach (parametric)

- Fit linear combinations of predictor variables so as to minimize the distance between the model’s predictions and the observed data.

  a) Pick the most efficient model by iteratively dropping predictor variable terms whose coefficients are least statistically significant and make the least overall contribution. (stepwise model selection)

  b) The prediction is the sum of values of the linear model over the subsequent variable values.

\[
Y = a + \sum_{i=1}^{n} b_i x_i
\]
Visualization: regression retirement model

PrRET(AGE,SVC) stepwise linear regression model
Classification approach (non-parametric)

View the data as two classes:
1. Actives that experience an event within time interval, and
2. Actives that do not experience the event.

a) Build a decision tree by successive partitioning that best allows us to classify each individual using a splitting criterion that either increases node class purity or maximizes information.

b) Prediction is the sum of the probabilities that an event will be experienced based on subsequent data values.

Advantage: Partitions can produce an understandable decision tree model which relates the prediction to patterns in the data.

Disadvantage: Because errors propagate down a hierarchical structure, the variability of tree predictions is large leading to an apparent lack of reproducibility and hence confidence in predictions.
Visualization: C&R tree retirement model
Divide retirement data into two collections: FY97-02 (chop); FY02-06 (calm)

LLNL Career Indefinite Retirement & Termination Rates

- Retirement
- Termination
- Attrition

Fiscal Year

Classification methods must learn temporal patterns also

- Temporal prediction vs. classification
  - There are classification models $M_y$ for each year $y$.
  - Each model produces insight into event correlated groupings.
  - If we amalgamate multiple contiguous years $\{y_1, \ldots, y_n\}$ and apply a partitioning technique a temporally sensitive variable should be included to allow the model to "learn" both how to group the data but also how these groups change temporally.

- Class size imbalance
  - In addition, tree methods work best when the classes are roughly in balance.
  - But in the case of attrition the events are scarce relative to the aggregate active population.
  - We can move from an imbalance of 1:30 to 1:12 by restricting our active population to employees qualifying for the retirement annuity.
Design of comparison experiment

• Livermore data subset (chop + calm) consists of 10 snapshots for FY97 through FY06.

• Our experiment consists of
  – Iteratively group these snapshots as singletons through 7-zes
  – At each training set aggregation level build models using 5 methods:
    • Stepwise linear regression (using AIC selection)
    • Stepwise logistic regression (using AIC selection)
    • Random forest
    • CnRT tree (uniform complexity parameter, not pruned)
    • Deviance tree (without cross-validation pruning)
  – Computing relative error in predicting subsequent year retirement
  – Computing mean and standard deviation of relative errors
### Detail for 5 year learning windows

<table>
<thead>
<tr>
<th>using retirement qualified</th>
<th>stepwise</th>
<th>forest</th>
<th>trees</th>
<th>actual</th>
<th>stepwise error</th>
<th>forest</th>
<th>tree rel. error</th>
</tr>
</thead>
<tbody>
<tr>
<td>train (w/ JP poly) yrs predict</td>
<td>linReg logReg</td>
<td>RF₄</td>
<td>CnRT</td>
<td>dTree</td>
<td>linReg logReg</td>
<td>RF₄</td>
<td>CnRT</td>
</tr>
<tr>
<td>1997-2001 5 2002</td>
<td>226 228</td>
<td>238 219 232</td>
<td>208</td>
<td>8.7% 9.6%</td>
<td>14.4% 5.3% 11.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998-2002 5 2003</td>
<td>235 241</td>
<td>245 274 251</td>
<td>233</td>
<td>0.9% 3.4%</td>
<td>5.2% 17.6% 7.7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999-2003 5 2004</td>
<td>262 260</td>
<td>266 258 237</td>
<td>258</td>
<td>1.6% 0.8%</td>
<td>3.1% 0.0% -8.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2000-2004 5 2005</td>
<td>284 279</td>
<td>274 261 283</td>
<td>263</td>
<td>8.0% 6.1%</td>
<td>4.2% -0.8% 7.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001-2005 5 2006</td>
<td>292 289</td>
<td>282 292 265</td>
<td>322</td>
<td>-9.3% -10.2%</td>
<td>-12.4% -9.3% -17.7%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Mean**
- 1.9% 1.9%
- 2.9% 2.6% 0.2%

**Std. Dev.**
- 0.072 0.076
- 0.097 0.099 0.126

**Chop**
- 8.7% 9.6%
- 14.4% 5.3% 11.5%

**Calm**
- 0.3% 0.0%
- 0.0% 1.9% -2.6%

- 0.072 0.072 0.083 0.113 0.125

- The exogenously perturbed period FY97-FY01 (internet boom/bust) impacts learning less as training window moves forward in time.
- Note yellow background highlighted mean and standard deviation.
### Detail for 3 year learning windows

<table>
<thead>
<tr>
<th>train (w/ JP poly)</th>
<th>yrs</th>
<th>predict</th>
<th>stepwise</th>
<th>[actual]</th>
<th>stepwise error</th>
<th>forest</th>
<th>tree rel. error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>linReg</td>
<td>logReg</td>
<td>linReg</td>
<td>logReg</td>
<td>RF4</td>
</tr>
<tr>
<td>1997-1999</td>
<td>3</td>
<td>2000</td>
<td>164</td>
<td>170</td>
<td>187</td>
<td>-12.3%</td>
<td>-9.1%</td>
</tr>
<tr>
<td>1998-2000</td>
<td>3</td>
<td>2001</td>
<td>187</td>
<td>193</td>
<td>201</td>
<td>-36.6%</td>
<td>-34.6%</td>
</tr>
<tr>
<td>1999-2001</td>
<td>3</td>
<td>2002</td>
<td>234</td>
<td>234</td>
<td>243</td>
<td>12.5%</td>
<td>12.5%</td>
</tr>
<tr>
<td>2000-2002</td>
<td>3</td>
<td>2003</td>
<td>255</td>
<td>255</td>
<td>261</td>
<td>9.4%</td>
<td>9.4%</td>
</tr>
<tr>
<td>2001-2003</td>
<td>3</td>
<td>2004</td>
<td>277</td>
<td>271</td>
<td>276</td>
<td>7.4%</td>
<td>5.0%</td>
</tr>
<tr>
<td>2002-2004</td>
<td>3</td>
<td>2005</td>
<td>267</td>
<td>259</td>
<td>260</td>
<td>1.5%</td>
<td>-1.5%</td>
</tr>
<tr>
<td>2003-2005</td>
<td>3</td>
<td>2006</td>
<td>277</td>
<td>275</td>
<td>275</td>
<td>-14.0%</td>
<td>-14.6%</td>
</tr>
</tbody>
</table>

| mean | -6.2% | -8.1% | -7.9% | -5.5% | -11.6% |
| std.dev | 0.175 | 0.164 | 0.168 | 0.178 | 0.175 |

- **A shorter learning window illustrates how the classification methods fare when retirement experiences rapid reversals.**
LLNL retirement qualified 5 year learning

Area under ROC curve: linReg = 0.81; ranForest = 0.80
Illustration showing why LANL chose 5 year learning windows for regression modeling

Finding right regression window

Vertical blue lines are sample standard deviations of prediction errors.

Purple discs are sample means of prediction errors.

Number of FYs together

1 2 3 4 5 6 7

-0.3 -0.2 -0.1 0.0 0.1 0.2 0.3

mean w/ stdv limits
Retirement prediction generalizations

- Cubic terms from Poly$^3$(AGE,SVC) are always part of both leaps or stepwise regression models and are of high importance in random forest models.

- Linear regression is slightly more accurate than logistic regression while the later has consistently less variance.

- A 5 year learning window is an empirical threshold for variance minimization when using regression.

- With a 5 year learning window logistic regression and random forests give nearly identical results.

- For a 3 year learning window tree methods are more accurate than regression resisting the effects of choppy data. Regression excels on calm data.
References and software
References

Workforce modeling


Regression


References, cont.

Tree classification

Hastie, Tibshirani, Friedman (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer


Random forest classification


Software used

- **Regression**
  - IBM SPSS 20 Statistics w/ Regression and Decision Trees
  - R 2.15 with package
    - leaps 2.9

- **Classification**
  - JMP Pro 10
  - R 2.15 with packages:
    - tree 1.0-23
    - rpart 4.1-0
    - randomForest 4.6-7
    - ada 2.0-3
    - ROCR 1.0-2
Wrap-up conclusions

• LANL’s regression experience was independently verified using LLNL data using both automated regression and tree based classification.
Supplemental Information

\[ D = \sum_i D_i \]

\[ D_i = -2 \sum_k n_{ik} \log p_{ik} \]
Each appointment type has its own characteristic attrition rate.
Joint Termination distribution from Frees study

<table>
<thead>
<tr>
<th>Age Nearest Birthday</th>
<th>Service &lt; 2</th>
<th>Service = 2, 3, 4</th>
<th>Service = 5-9</th>
<th>Service ≥ 10</th>
<th>Total</th>
<th>Service &lt; 2</th>
<th>Service = 2, 3, 4</th>
<th>Service = 5-9</th>
<th>Service ≥ 10</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>552</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>574</td>
<td>30.62</td>
<td>4.55</td>
<td>0.00</td>
<td>0.00</td>
<td>29.62</td>
</tr>
<tr>
<td>19</td>
<td>1,521</td>
<td>109</td>
<td>0</td>
<td>0</td>
<td>1,630</td>
<td>27.42</td>
<td>7.34</td>
<td>0.00</td>
<td>0.00</td>
<td>26.07</td>
</tr>
<tr>
<td>20</td>
<td>3,962</td>
<td>517</td>
<td>2</td>
<td>0</td>
<td>4,581</td>
<td>16.58</td>
<td>7.52</td>
<td>0.00</td>
<td>0.00</td>
<td>15.37</td>
</tr>
<tr>
<td>21</td>
<td>8,001</td>
<td>3,466</td>
<td>25</td>
<td>0</td>
<td>11,494</td>
<td>22.47</td>
<td>18.90</td>
<td>4.00</td>
<td>0.00</td>
<td>21.32</td>
</tr>
<tr>
<td>22</td>
<td>11,144</td>
<td>7,867</td>
<td>166</td>
<td>0</td>
<td>19,167</td>
<td>24.10</td>
<td>21.17</td>
<td>14.46</td>
<td>0.00</td>
<td>22.82</td>
</tr>
<tr>
<td>23</td>
<td>13,142</td>
<td>10,766</td>
<td>579</td>
<td>0</td>
<td>24,487</td>
<td>22.68</td>
<td>19.69</td>
<td>16.23</td>
<td>0.00</td>
<td>21.22</td>
</tr>
<tr>
<td>24</td>
<td>14,735</td>
<td>13,305</td>
<td>2,138</td>
<td>0</td>
<td>30,178</td>
<td>22.16</td>
<td>17.71</td>
<td>14.31</td>
<td>0.00</td>
<td>19.64</td>
</tr>
<tr>
<td>25</td>
<td>14,679</td>
<td>16,119</td>
<td>4,686</td>
<td>2</td>
<td>35,486</td>
<td>20.72</td>
<td>17.19</td>
<td>13.72</td>
<td>0.00</td>
<td>18.19</td>
</tr>
<tr>
<td>26</td>
<td>14,206</td>
<td>17,942</td>
<td>7,473</td>
<td>17</td>
<td>39,636</td>
<td>20.34</td>
<td>16.40</td>
<td>11.07</td>
<td>23.53</td>
<td>16.81</td>
</tr>
<tr>
<td>27</td>
<td>13,320</td>
<td>18,837</td>
<td>10,784</td>
<td>136</td>
<td>43,077</td>
<td>19.75</td>
<td>15.31</td>
<td>9.75</td>
<td>10.29</td>
<td>15.27</td>
</tr>
<tr>
<td>28</td>
<td>12,815</td>
<td>18,444</td>
<td>14,604</td>
<td>421</td>
<td>46,284</td>
<td>18.94</td>
<td>14.44</td>
<td>9.25</td>
<td>9.50</td>
<td>14.00</td>
</tr>
<tr>
<td>29</td>
<td>11,973</td>
<td>17,666</td>
<td>17,483</td>
<td>1,600</td>
<td>48,744</td>
<td>17.92</td>
<td>13.97</td>
<td>8.67</td>
<td>7.88</td>
<td>12.80</td>
</tr>
<tr>
<td>30</td>
<td>11,289</td>
<td>16,642</td>
<td>19,466</td>
<td>3,442</td>
<td>50,829</td>
<td>18.38</td>
<td>13.64</td>
<td>8.45</td>
<td>6.22</td>
<td>12.21</td>
</tr>
<tr>
<td>31</td>
<td>10,804</td>
<td>15,994</td>
<td>21,026</td>
<td>5,678</td>
<td>53,502</td>
<td>18.19</td>
<td>12.99</td>
<td>7.92</td>
<td>5.13</td>
<td>11.21</td>
</tr>
<tr>
<td>32</td>
<td>10,640</td>
<td>15,684</td>
<td>22,161</td>
<td>8,659</td>
<td>57,144</td>
<td>17.73</td>
<td>12.69</td>
<td>7.83</td>
<td>5.49</td>
<td>10.65</td>
</tr>
<tr>
<td>33</td>
<td>10,422</td>
<td>15,456</td>
<td>22,237</td>
<td>12,398</td>
<td>60,512</td>
<td>16.75</td>
<td>12.05</td>
<td>7.51</td>
<td>5.32</td>
<td>9.81</td>
</tr>
<tr>
<td>34</td>
<td>9,894</td>
<td>15,073</td>
<td>22,117</td>
<td>16,100</td>
<td>63,184</td>
<td>16.18</td>
<td>11.23</td>
<td>7.39</td>
<td>5.12</td>
<td>9.11</td>
</tr>
<tr>
<td>35</td>
<td>9,617</td>
<td>14,505</td>
<td>21,650</td>
<td>19,649</td>
<td>65,421</td>
<td>16.29</td>
<td>10.95</td>
<td>7.21</td>
<td>5.01</td>
<td>8.72</td>
</tr>
<tr>
<td>36</td>
<td>9,084</td>
<td>13,934</td>
<td>20,859</td>
<td>22,883</td>
<td>66,760</td>
<td>16.15</td>
<td>10.94</td>
<td>6.75</td>
<td>4.87</td>
<td>8.26</td>
</tr>
<tr>
<td>37</td>
<td>8,518</td>
<td>13,381</td>
<td>19,798</td>
<td>25,385</td>
<td>67,082</td>
<td>15.47</td>
<td>11.22</td>
<td>6.73</td>
<td>4.76</td>
<td>7.99</td>
</tr>
<tr>
<td>38</td>
<td>8,002</td>
<td>12,502</td>
<td>18,967</td>
<td>27,161</td>
<td>66,642</td>
<td>15.01</td>
<td>10.53</td>
<td>6.38</td>
<td>4.30</td>
<td>7.49</td>
</tr>
<tr>
<td>40</td>
<td>7,233</td>
<td>10,789</td>
<td>16,949</td>
<td>29,425</td>
<td>64,396</td>
<td>15.82</td>
<td>10.31</td>
<td>5.90</td>
<td>4.19</td>
<td>6.97</td>
</tr>
<tr>
<td>41</td>
<td>6,803</td>
<td>10,254</td>
<td>15,976</td>
<td>29,783</td>
<td>62,816</td>
<td>15.54</td>
<td>10.12</td>
<td>5.91</td>
<td>3.86</td>
<td>6.67</td>
</tr>
<tr>
<td>42</td>
<td>6,166</td>
<td>9,527</td>
<td>14,817</td>
<td>29,521</td>
<td>60,031</td>
<td>15.44</td>
<td>9.55</td>
<td>5.93</td>
<td>3.88</td>
<td>6.47</td>
</tr>
<tr>
<td>43</td>
<td>5,666</td>
<td>8,739</td>
<td>13,562</td>
<td>28,945</td>
<td>57,132</td>
<td>15.55</td>
<td>9.83</td>
<td>5.63</td>
<td>3.81</td>
<td>6.37</td>
</tr>
<tr>
<td>44</td>
<td>5,427</td>
<td>8,058</td>
<td>12,803</td>
<td>28,045</td>
<td>54,335</td>
<td>15.16</td>
<td>9.58</td>
<td>5.79</td>
<td>3.79</td>
<td>6.26</td>
</tr>
<tr>
<td>45</td>
<td>5,083</td>
<td>7,634</td>
<td>12,015</td>
<td>27,074</td>
<td>51,806</td>
<td>14.89</td>
<td>9.42</td>
<td>5.84</td>
<td>3.80</td>
<td>6.19</td>
</tr>
<tr>
<td>46</td>
<td>4,453</td>
<td>6,964</td>
<td>11,066</td>
<td>25,720</td>
<td>48,223</td>
<td>15.14</td>
<td>9.71</td>
<td>5.93</td>
<td>3.55</td>
<td>6.06</td>
</tr>
<tr>
<td>47</td>
<td>4,190</td>
<td>6,290</td>
<td>10,321</td>
<td>24,736</td>
<td>45,537</td>
<td>14.89</td>
<td>9.22</td>
<td>5.47</td>
<td>3.65</td>
<td>5.87</td>
</tr>
<tr>
<td>48</td>
<td>4,026</td>
<td>5,880</td>
<td>9,791</td>
<td>24,055</td>
<td>43,752</td>
<td>14.48</td>
<td>9.73</td>
<td>5.42</td>
<td>3.75</td>
<td>5.92</td>
</tr>
</tbody>
</table>
**Accuracy of method over time**
(Has the method predicted the next year’s attrition?)

**Significant events:**
- 2004 – safety issues and stand down
- 2005 – contract negotiations
- 2006 – contract transition
- 2007 – first full year under new management
- 2008 – Self Selection Program
- 2009 – 2011- Economy
- 2012 – Voluntary Separation Program

![Graph showing actual vs projected attrition from FY 2002 to FY 2013](image-url)
## Individual tree pruning

<table>
<thead>
<tr>
<th>LLNL retirement qualified</th>
<th>stepwise</th>
<th>forest</th>
<th>pruned trees</th>
<th>best tuned</th>
<th>actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>train (w/ JP poly) yrs predict linReg logReg</td>
<td></td>
<td>RF₄</td>
<td>CnRT SPSS JMP dTree</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1997-2001 5 2002</td>
<td>226</td>
<td>228</td>
<td>238</td>
<td>218 217</td>
<td>206 206</td>
</tr>
<tr>
<td>1998-2002 5 2003</td>
<td>235</td>
<td>241</td>
<td>245</td>
<td>242 229</td>
<td>212 231</td>
</tr>
<tr>
<td>2000-2004 5 2005</td>
<td>284</td>
<td>279</td>
<td>274</td>
<td>268 253</td>
<td>259 261</td>
</tr>
</tbody>
</table>

- **auc ROC** = 0.813 0.801

CART - Gini splitting w/ pruning by **resub 1SE**
Deviance splitting w/ 10-fold cross-validation stop **known**

- CnRT `rpart` using resubstitution misclassification estimate to prune
- SPSS Decision Tree package using 1SE pruning condition (same start as `rpart`)
- JMP uses deviance splitting stopped by 10-fold cross-validation misclass estimate
- dTree `prune.tree` finding best prediction accuracy size tree by linear search
Pruning shows best possible

<table>
<thead>
<tr>
<th>LLNL retirement qualified</th>
<th>stepwise error</th>
<th>forest</th>
<th>tree rel. error = (eR-aR)/aR</th>
</tr>
</thead>
<tbody>
<tr>
<td>train (w/ JP poly) yrs predict</td>
<td>actual</td>
<td>linReg</td>
<td>logReg</td>
</tr>
<tr>
<td>1997-2001 5 2002</td>
<td>208</td>
<td>8.7%</td>
<td>9.6%</td>
</tr>
<tr>
<td>1998-2002 5 2003</td>
<td>233</td>
<td>0.9%</td>
<td>3.4%</td>
</tr>
<tr>
<td>1999-2003 5 2004</td>
<td>258</td>
<td>1.6%</td>
<td>0.8%</td>
</tr>
<tr>
<td>2000-2004 5 2005</td>
<td>263</td>
<td>8.0%</td>
<td>6.1%</td>
</tr>
<tr>
<td>2001-2005 5 2006</td>
<td>322</td>
<td>-9.3%</td>
<td>-10.2%</td>
</tr>
</tbody>
</table>

mean 1.9% | 1.9% | 2.9% | 0.6% | -4.9% | -7.2% | -4.5% |

std.dev 0.072 | 0.076 | 0.097 | 0.053 | 0.083 | 0.062 | 0.065 |

chop 8.7% | 9.6% | 14.4% | 4.8% | 4.3% | -1.0% | -1.0% |

calm 0.3% | 0.0% | 0.0% | -0.5% | -7.2% | -8.8% | -5.3% |

0.072 | 0.072 | 0.083 | 0.054 | 0.075 | 0.059 | 0.072 |

- **CnRT rpart** using resubstitution misclassification estimate to prune
- **SPSS Decision Tree package** using 1SE pruning condition (same start as **rpart**)
- **JMP** uses deviance splitting stopped by 10-fold cross-validation misclass estimate
- **dTree prune.tree** finding best prediction accuracy size tree by linear search
JMP interactive partitioning reveals different predictor patterns for termination vs. the retirement models.