# Sensible Approaches to Handling Unbalanced Data"

Norm Matloff
University of California at Davis

Bay Area R Users Group
September 8, 2020

URL for these slides (repeated on final slide):
http://heather.cs.ucdavis.edu/BARUGunbal.pdf

# Overview

# Overview

- Classification problem, multiple classes.

# Overview

- Classification problem, multiple classes.
- $n_i$ = number of training data points in class $i$

# Overview

- Classification problem, multiple classes.
- $n_i$ = number of training data points in class $i$
- Termed *unbalanced* if some $n_i$ is highly dominant, say 100X larger than the others.

# Overview

- Classification problem, multiple classes.
- $n_i$ = number of training data points in class $i$
- Termed *unbalanced* if some $n_i$ is highly dominant, say 100X larger than the others.
- "Problem": Standard modeling techniques will tend to predict all, or almost all, new cases to be the dominant class.

# Overview

- Classification problem, multiple classes.
- $n_i$ = number of training data points in class $i$
- Termed *unbalanced* if some $n_i$ is highly dominant, say 100X larger than the others.
- "Problem": Standard modeling techniques will tend to predict all, or almost all, new cases to be the dominant class.
- Standard "solution": Artificially balance the training data classes via resampling.

# Overview

- Classification problem, multiple classes.
- $n_i$ = number of training data points in class $i$
- Termed *unbalanced* if some $n_i$ is highly dominant, say 100X larger than the others.
- "Problem": Standard modeling techniques will tend to predict all, or almost all, new cases to be the dominant class.
- Standard "solution": Artificially balance the training data classes via resampling.
- BUT NOT A GOOD IDEA. Distortionary and harmful.

# Overview

- Classification problem, multiple classes.
- $n_i$ = number of training data points in class $i$
- Termed *unbalanced* if some $n_i$ is highly dominant, say 100X larger than the others.
- "Problem": Standard modeling techniques will tend to predict all, or almost all, new cases to be the dominant class.
- Standard "solution": Artificially balance the training data classes via resampling.
- BUT NOT A GOOD IDEA. Distortionary and harmful.
- One can do much better.

# Preparation

# Preparation

"Read directions before use"

# Preparation

"Read directions before use"

- Data Science is NOT Computer Science.

# Preparation

"Read directions before use"

- Data Science is NOT Computer Science.
- DS is NOT just a matter of knowing a bunch of packages and functions.

# Preparation

"Read directions before use"

- Data Science is NOT Computer Science.
- DS is NOT just a matter of knowing a bunch of packages and functions.
- Good DS means:

# Preparation

"Read directions before use"

- Data Science is NOT Computer Science.
- DS is NOT just a matter of knowing a bunch of packages and functions.
- Good DS means:
  - Careful thought about one's goals.
  - Careful selection of functions (or writing new code altogether) to fit those goals.
  - Thoughtful interpretation of one's results, possibly modifying and re-running.

# Preparation

"Read directions before use"

- Data Science is NOT Computer Science.
- DS is NOT just a matter of knowing a bunch of packages and functions.
- Good DS means:
  - Careful thought about one's goals.
  - Careful selection of functions (or writing new code altogether) to fit those goals.
  - Thoughtful interpretation of one's results, possibly modifying and re-running.
- Beware of complicated solutions to simple problems.

Norm Matloff
University of California at Davis

# Provenance of This Talk

# Provenance of This Talk

- N. Matloff, *Statistical Regression and Classification: from Linear Models to Machine Learning*, CRC, 2017 (recipient of the Ziegal Award), 193-202
- John Mount, *Learning from Imbalanced Classes*, *https://win-vector.com/2020/08/07/dont-use-classification-rules-for-classification-problems/*, 2020
- More recent joint work with John Mount and Nina Zumel.

# Motivating Example: Missed Appointments Data

# Motivating Example: Missed Appointments Data

- Want to predict no-shows for medical appointments.

# Motivating Example: Missed Appointments Data

- Want to predict no-shows for medical appointments.
- About 20% of training data is no-shows.

# Motivating Example: Missed Appointments Data

- Want to predict no-shows for medical appointments.
- About 20% of training data is no-shows.
- Try, say, k-NN (from **regtools** package).

# Motivating Example: Missed Appointments Data

- Want to predict no-shows for medical appointments.
- About 20% of training data is no-shows.
- Try, say, k-NN (from **regtools** package).

```
> preds ←
    kNN(ma2[,-89],ma2[,89],ma2[idxs,-89],50)
> table(preds$ypreds)
  0    1
 53 9947
```

j

- Almost all predictions are for Class 1, not very useful.

# Motivating Example: Missed Appointments Data

- Want to predict no-shows for medical appointments.
- About 20% of training data is no-shows.
- Try, say, k-NN (from **regtools** package).

```
> preds ←
    kNN(ma2[,−89],ma2[,89],ma2[idxs,−89],50)
> table(preds$ypreds)
  0     1
 53  9947
```

j

- Almost all predictions are for Class 1, not very useful. (There is also a question of quality of fit. A local-linear model might be better, not pursued here.)

Norm Matloff
University of California at Davis

# The "Follow the Crowd" Approach

Norm Matloff
University of California at Davis

# The "Follow the Crowd" Approach

- Source of the problem: Unbalanced data.

# The "Follow the Crowd" Approach

- Source of the problem: Unbalanced data.
- Assumed solution: Force the data to be balanced, by resampling.

# The "Follow the Crowd" Approach

- Source of the problem: Unbalanced data.
- Assumed solution: Force the data to be balanced, by resampling.
    - Downsample: Throw out data from dominant class.
    - Upsample: Make up extra data for minority class.
    - Resample: Essentially a bootstrap sampling, but weighted toward the minority class.

# No Justification for Such Approaches

# No Justification for Such Approaches

- None of those approaches makes sense.

# No Justification for Such Approaches

- None of those approaches makes sense.
- Throw OUT data? Really?

## No Justification for Such Approaches

- None of those approaches makes sense.
- Throw OUT data? Really?
- Distort the data? Has anyone thought about the consequences?

Norm Matloff
University of California at Davis

# No Justification for Such Approaches

- None of those approaches makes sense.
- Throw OUT data? Really?
- Distort the data? Has anyone thought about the consequences?
- And anyway, what's wrong with the simple, obvious "person on the street" solution?

# Person-on-the-Street Approach

# Person-on-the-Street Approach

- Those with no background may have more common sense.

# Person-on-the-Street Approach

- Those with no background may have more common sense.
- Person-on-street would say, "Well, just identify which patients are at substantial risk of being no-shows."

# Person-on-Street, cont'd.

# Person-on-Street, cont'd.

Fitted values, training data:

# Person-on-Street, cont'd.

Fitted values, training data:

```
> table(preds$regests)
0.34  0.4 0.42 0.44 0.46 0.48  0.5 0.52 0.54 0.56
   5    6    2    7   10    9   14   26   39   47
0.58  0.6 0.62 0.64 0.66 0.68  0.7 0.72 0.74 0.76
  89  143  156  205  273  343  340  480  585  631
0.78  0.8 0.82 0.84 0.86 0.88  0.9 0.92 0.94 0.96
 757  840  861  901  847  778  621  437  285  153
0.98    1
  62   48
```

# Person-on-Street, cont'd.

Fitted values, training data:

```
> table(preds$regests)
0.34   0.4  0.42  0.44  0.46  0.48   0.5  0.52  0.54  0.56
   5     6     2     7    10     9    14    26    39    47
0.58   0.6  0.62  0.64  0.66  0.68   0.7  0.72  0.74  0.76
  89   143   156   205   273   343   340   480   585   631
0.78   0.8  0.82  0.84  0.86  0.88   0.9  0.92  0.94  0.96
 757   840   861   901   847   778   621   437   285   153
0.98     1
  62    48
```

E.g. 2779 have risk $\geq$ 0.25 of no-show.

# Person-on-Street, cont'd.

Fitted values, training data:

```
> table(preds$regests)
0.34  0.4 0.42 0.44 0.46 0.48  0.5 0.52 0.54 0.56
   5    6    2    7   10    9   14   26   39   47
0.58  0.6 0.62 0.64 0.66 0.68  0.7 0.72 0.74 0.76
  89  143  156  205  273  343  340  480  585  631
0.78  0.8 0.82 0.84 0.86 0.88  0.9 0.92 0.94 0.96
 757  840  861  901  847  778  621  437  285  153
0.98    1
  62   48
```

E.g. 2779 have risk $\geq 0.25$ of no-show.
So, just flag future cases with risk over 0.25, and give them extra reminders about the appointment etc.

Norm Matloff
University of California at Davis

# Comments

# Comments

- The fit, without balancing, is not "wrong."

## Comments

- The fit, without balancing, is not "wrong." The use to which it was put was not useful, true, but not wrong.

# Comments

- The fit, without balancing, is not "wrong." The use to which it was put was not useful, true, but not wrong.
- But in predicting the class of new case, this assumed goal is min overall misclassification rate,

# Comments

- The fit, without balancing, is not "wrong." The use to which it was put was not useful, true, but not wrong.
- But in predicting the class of new case, this assumed goal is min overall misclassification rate, again not useful here.

# Comments

- The fit, without balancing, is not "wrong." The use to which it was put was not useful, true, but not wrong.
- But in predicting the class of new case, this assumed goal is min overall misclassification rate, again not useful here.
- *Could* do a formal utility analysis, different costs for different types of misclassification.

## Comments

- The fit, without balancing, is not "wrong." The use to which it was put was not useful, true, but not wrong.
- But in predicting the class of new case, this assumed goal is min overall misclassification rate, again not useful here.
- *Could* do a formal utility analysis, different costs for different types of misclassification. Good if we want to impress people with our math prowess.

# Comments

- The fit, without balancing, is not "wrong." The use to which it was put was not useful, true, but not wrong.
- But in predicting the class of new case, this assumed goal is min overall misclassification rate, again not useful here.
- *Could* do a formal utility analysis, different costs for different types of misclassification. Good if we want to impress people with our math prowess.
- But the person-on-the-street approach is simpler and fulfills our goals.

# Comments

- The fit, without balancing, is not "wrong." The use to which it was put was not useful, true, but not wrong.
- But in predicting the class of new case, this assumed goal is min overall misclassification rate, again not useful here.
- *Could* do a formal utility analysis, different costs for different types of misclassification. Good if we want to impress people with our math prowess.
- But the person-on-the-street approach is simpler and fulfills our goals.
- And, analysis with artificially balanced data IS wrong. (Next slide.)

# Harmful Distortions

# Harmful Distortions

- Let $\pi_i$ denote the true population proportion for class $i$.

# Harmful Distortions

- Let $\pi_i$ denote the true population proportion for class $i$. E.g. in the Missed Appointments example, $\pi_1$ is the population proportion of patients who are no-shows.

# Harmful Distortions

- Let $\pi_i$ denote the true population proportion for class $i$. E.g. in the Missed Appointments example, $\pi_1$ is the population proportion of patients who are no-shows.
- The algorithm you use, doesn't matter which, implicitly assumes that the class sizes $n_i$ reflect the $\pi_i$, i.e. $\widehat{\pi}_i = n_i / n$.

# Harmful Distortions

- Let $\pi_i$ denote the true population proportion for class $i$. E.g. in the Missed Appointments example, $\pi_1$ is the population proportion of patients who are no-shows.
- The algorithm you use, doesn't matter which, implicitly assumes that the class sizes $n_i$ reflect the $\pi_i$, i.e. $\widehat{\pi}_i = n_i/n$.
- So, if you artificially balance your data, your algorithm will think all the $\pi_i$ are equal.

# Harmful Distortions

- Let $\pi_i$ denote the true population proportion for class $i$. E.g. in the Missed Appointments example, $\pi_1$ is the population proportion of patients who are no-shows.
- The algorithm you use, doesn't matter which, implicitly assumes that the class sizes $n_i$ reflect the $\pi_i$, i.e. $\widehat{\pi}_i = n_i / n$.
- So, if you artificially balance your data, your algorithm will think all the $\pi_i$ are equal.
- Thus, in predicting a new case, your algorithm will OVERestimate the (conditional) probability of a class for which $\pi_i$ is smaller than average, and UNDERestimate in the case of a class for which $\pi_i$ is larger than average,

# Harmful Distortions

- Let $\pi_i$ denote the true population proportion for class $i$. E.g. in the Missed Appointments example, $\pi_1$ is the population proportion of patients who are no-shows.
- The algorithm you use, doesn't matter which, implicitly assumes that the class sizes $n_i$ reflect the $\pi_i$, i.e. $\widehat{\pi}_i = n_i/n$.
- So, if you artificially balance your data, your algorithm will think all the $\pi_i$ are equal.
- Thus, in predicting a new case, your algorithm will OVERestimate the (conditional) probability of a class for which $\pi_i$ is smaller than average, and UNDERestimate in the case of a class for which $\pi_i$ is larger than average,
- So, YES, it MATTERS.

Norm Matloff
University of California at Davis

# What If Data Are Already Artificially Balanced?

Norm Matloff
University of California at Davis

## What If Data Are Already Artificially Balanced?

- We may have data, given to us by others, in which the actual data collection was done in a balanced manner.

## What If Data Are Already Artificially Balanced?

- We may have data, given to us by others, in which the actual data collection was done in a balanced manner.
- Ah, so no problem, right? The data is already balanced.

# What If Data Are Already Artificially Balanced?

- We may have data, given to us by others, in which the actual data collection was done in a balanced manner.
- Ah, so no problem, right? The data is already balanced. Wrong!

## What If Data Are Already Artificially Balanced?

- We may have data, given to us by others, in which the actual data collection was done in a balanced manner.
- Ah, so no problem, right? The data is already balanced. Wrong!
- Same problem as above, wrong estimates of the $\pi_i$.

## What If Data Are Already Artificially Balanced?

- We may have data, given to us by others, in which the actual data collection was done in a balanced manner.
- Ah, so no problem, right? The data is already balanced. Wrong!
- Same problem as above, wrong estimates of the $\pi_i$.
- SOLUTION: If you have estimates of the true class probabilities, I have an update formula to convert the estimated conditional class probabilities to the proper values.

## What If Data Are Already Artificially Balanced?

- We may have data, given to us by others, in which the actual data collection was done in a balanced manner.
- Ah, so no problem, right? The data is already balanced. Wrong!
- Same problem as above, wrong estimates of the $\pi_i$.
- SOLUTION: If you have estimates of the true class probabilities, I have an update formula to convert the estimated conditional class probabilities to the proper values. Derivation in *github.com/matloff/regtools/UnbalancedClasses.md*.

# What If Data Are Already Artificially Balanced?

- We may have data, given to us by others, in which the actual data collection was done in a balanced manner.
- Ah, so no problem, right? The data is already balanced. Wrong!
- Same problem as above, wrong estimates of the $\pi_i$.
- SOLUTION: If you have estimates of the true class probabilities, I have an update formula to convert the estimated conditional class probabilities to the proper values. Derivation in *github.com/matloff/regtools/UnbalancedClasses.md*.
- Example: UCI Letters data. All $n_i/n \approx 1/26$, but true values at *http://www.math.cornell.edu/ mec/2003-2004/cryptography/subs/frequencies.html*.

# What If Class Probabilities Vary by Site?

# What If Class Probabilities Vary by Site?

- Interesting example (Efron *JASA* $\rightarrow$ *Scientific American*).

# What If Class Probabilities Vary by Site?

- Interesting example (Efron *JASA* $\rightarrow$ *Scientific American*).
- Machine diagnosis of pneumonia from X-ray images, Mt. Sinai Hospital.

Norm Matloff
University of California at Davis

## What If Class Probabilities Vary by Site?

- Interesting example (Efron *JASA* → *Scientific American*).
- Machine diagnosis of pneumonia from X-ray images, Mt. Sinai Hospital.
- Predicted new cases at Mt. Sinai well,

# What If Class Probabilities Vary by Site?

- Interesting example (Efron *JASA* → *Scientific American*).
- Machine diagnosis of pneumonia from X-ray images, Mt. Sinai Hospital.
- Predicted new cases at Mt. Sinai well, but not at other facilities.

# What If Class Probabilities Vary by Site?

- Interesting example (Efron *JASA* → *Scientific American*).
- Machine diagnosis of pneumonia from X-ray images, Mt. Sinai Hospital.
- Predicted new cases at Mt. Sinai well, but not at other facilities.
- The researchers found cause: The $\pi_i$ vary from hospital to hospital.

# What If Class Probabilities Vary by Site?

- Interesting example (Efron *JASA* $\rightarrow$ *Scientific American*).
- Machine diagnosis of pneumonia from X-ray images, Mt. Sinai Hospital.
- Predicted new cases at Mt. Sinai well, but not at other facilities.
- The researchers found cause: The $\pi_i$ vary from hospital to hospital.
- This can be solved using my update formula.

Norm Matloff
University of California at Davis

# R Software

# R Software

Given a new case, how do we get those (conditional) estimated probabilities of the different classes?

## R Software

Given a new case, how do we get those (conditional) estimated probabilities of the different classes?

E.g. **glm()**, in the Missed Appointments data, on a set of new cases **ccf**:

```
> glout ← glm(Class ~ .,data=ccf,family=binomial)
> condprobs ← predict(glout,ccf,type='response')
> tocheck ← which(condprobs > 0.25)
> names(tocheck) ← NULL
> head(tocheck)
[1]  542 6109 6332 6335 6337 6339
```

# R Software

Given a new case, how do we get those (conditional) estimated probabilities of the different classes?
E.g. **glm()**, in the Missed Appointments data, on a set of new cases **ccf**:

```
> glout ← glm(Class ∼ .,data=ccf,family=binomial)
> condprobs ← predict(glout,ccf,type='response')
> tocheck ← which(condprobs > 0.25)
> names(tocheck) ← NULL
> head(tocheck)
[1]  542 6109 6332 6335 6337 6339
```

So we'd check cases 542, 6109 etc. by hand.

# R Software, cont'd.

# R Software, cont'd.

E.g. **randomForests()**.

## R Software, cont'd.

E.g. **randomForests()**.
More work to do here, but a wrapper could be written:

# R Software, cont'd.

E.g. **randomForests()**.

More work to do here, but a wrapper could be written:

```
> ccf$Class <- as.factor(ccf$Class)
> rfout <- randomForest(Class ~ ., data=ccf)
> predout <- predict(rfout,ccf,type='response')
> treeguesses <-
     predout$individual  # class guesses, each tree
> tgs <- as.matrix(treeguesses)
> probs <- apply(tgs,1,
     function(rw) mean(as.numeric(rw)))
> tocheck <- which(probs > 0.25)
> head(tocheck)
[1]   70  542  624 1747 4921 6109
```

# R Software, cont'd.

# R Software, cont'd.

Also, the formula mentioned earlier for updating from incorrect to correct unconditional class probabilities is implemented in the **regtools**:

```
classadjust(econdprobs,wrongprob1,trueprob1)
```

# R Software, cont'd.

Also, the formula mentioned earlier for updating from incorrect to correct unconditional class probabilities is implemented in the **regtools**:

```
classadjust(econdprobs,wrongprob1,trueprob1)
```

By the way:

# R Software, cont'd.

Also, the formula mentioned earlier for updating from incorrect to correct unconditional class probabilities is implemented in the **regtools**:

```
classadjust(econdprobs,wrongprob1,trueprob1)
```

By the way:
The **regtools** package has been greatly expanded since its last upload to CRAN.

# R Software, cont'd.

Also, the formula mentioned earlier for updating from incorrect to correct unconditional class probabilities is implemented in the **regtools**:

```
classadjust(econdprobs,wrongprob1,trueprob1)
```

By the way:
The **regtools** package has been greatly expanded since its last upload to CRAN.
Now more than 80 functions for regression, classification and machine learning.