A MULTI-MODEL DATA MINING EXPLORATION OF U.S. MASS KILLINGS

Benjamin Heindl, Qi Luan, Christopher Pate, Daniel Serna
IST707 – Project Presentation, 14 June 2023

AGENDA

- Introduction
- Data Source; Overview
- Key Analytic Question(s) - KAQs
- Methods and Results
- Key Takeaways

INTRODUCTION

The subject of U.S. mass killings

Questions are often focused on the characteristics of these tragic events

Publicly available data are available to help answer key questions; data also include narrative summaries of killing incidents

Purpose of this work is to (1) leverage publicly available data on U.S. mass shootings and (2) use EDA and range of data mining methods to analyze the data, and (3) answer key analytic question(s)

Data source: Mass Killings in America, 2006 - Present, source: https://data.world/associatedpress/mass-killings-public

Owner: Associated Press-USA TODAY-Northeastern University

Methodology: Supplementary Homicide Reports (SHR; FBI), case verification through media accounts, court documents, journal articles, books and local law enforcement records obtained through FOIA requests

DATA SOURCE - OVERVIEW

LIST OF PUBLIC VARIABLES

Incident	*Offender*	*Victim*	*Weapon*
incident_id	incident_id	incident_id	incident_id
date	offender_id	victim_id	weapon_id
city	firstname	age	weapon_type
state	middlename	race	gun_class
num_offenders	lastname	sex	gun_type
num_victims_killed	suffix	vorelationship	
num_victims_injured	age		
firstcod	race		
secondcod	sex		
type	suicide		
situation_type	deathcause		
location_type	outcome		
location_type	criminal_justice_process		
longitude	sentence_type		
latitude	sentence_details		
GPS_point			
narrative			

DATA SOURCE - OVERVIEW

Q1: How can narrative summaries within the mass killings dataset be transformed into a tokenized structure?

Q2: Using the TF-IDF weighted data structure, how well do machine learning methods perform in classifying the type of killing?

Q3: Of the models considered in the analysis, which models perform best?

Q4: How does rebalancing the multi-class outcome (“type”) change model performance?

KEY ANALYTIC QUESTION(S) - KAQS

ANALYTIC PROCESS – FOLLOWING THE ANALYTIC LIFECYCLE

Business understanding

- Background, business objectives, overall project plan – integrate project requirements with coursework
- Use of R, multiple packages (notably *caret*)

Data understanding

- Collection of data, initial exploratory data analysis (EDA)
- Data quality – identify missing values, iterate throughout all phases

Data preparation

- Data cleansing – missing values, analysis and removal as appropriate
- Central focus involved text mining methods: creation of a document-term matrix from narrative elements, creation of appropriate data structure and integration with the core dataset to support the modeling process

Modeling

- Classification trees, recursive partitioning, C5, support vector machines, random forest and *k*-means clustering

EXPLORATORY DATA ANALYSIS

Figure 1. *Data Summary – Incidents Dataset*

	Descriptions	Value
1	Sample size (nrow)	553
2	No. of variables (ncol)	17
3	No. of numeric/interger variables	7
4	No. of factor variables	0
5	No. of text variables	10
6	No. of logical variables	0
7	No. of identifier variables	5
8	No. of date variables	0
9	No. of zero variance variables (uniform)	0
10	%. of variables having complete cases	88.24% (15)
11	%. of variables having >0% and <50% missing cases	5.88% (1)
12	%. of variables having >=50% and <90% missing cases	5.88% (1)
13	%. of variables having >=90% missing cases	0% (0)

Figure 2. *Data Summary – Document-Term Matrix, Incidents Dataset*

	Descriptions	Value
1	Sample size (nrow)	553
2	No. of variables (ncol)	168
3	No. of numeric/interger variables	168
4	No. of factor variables	0
5	No. of text variables	0
6	No. of logical variables	0
7	No. of identifier variables	1
8	No. of date variables	0
9	No. of zero variance variables (uniform)	0
10	%. of variables having complete cases	100% (168)
11	%. of variables having >0% and <50% missing cases	0% (0)
12	%. of variables having >=50% and <90% missing cases	0% (0)
13	%. of variables having >=90% missing cases	0% (0)

TEXT ANALYSIS: PREPROCESSING

Figure 3. *Code Snippet: Text Analysis Preprocessing*

```
## create corpus; convert the narrative field within the incidents dataset to the corpus; conduct the following preprocessing steps
## 1. convert all text to lower case, 2. remove punctuation, and 3. remove stopwords
incident_corpus <- Corpus(VectorSource(incidents2$narrative))
incident_corpus <- tm_map(incident_corpus, PlainTextDocument)
incident_corpus <- tm_map(incident_corpus, tolower)
incident_corpus <- tm_map(incident_corpus, removePunctuation)
incident_corpus <- tm_map(incident_corpus, removeWords, stopwords("english"))

## Two distinct approaches here - TF and TFIDF weighting; first is TF
## create the document term matrix w TF weighting; extract frequently occurring words (target roughly 170 words for the analysis)
## create dataframe of sparse matrix, one word per column
dtm <- DocumentTermMatrix(incident_corpus)
notSparse <- removeSparseTerms(dtm, 0.975)
finalWords <- as.data.frame(as.matrix(notSparse), stringsAsFactors = FALSE)
head (finalWords)

## create index column; check dimensions of data frame; view column names; examine subset and check summary of one of the terms
## conduct EDA
finalWords2 <- cbind(index = 1:nrow(finalWords), finalWords)
dim (finalWords2)
ExpData(finalWords2, type=1)
ExpData(finalWords2, type=2)
colnames(finalWords2)
finalWords2[148:158, 26:29]
summary (finalWords2$fire)
```

Figure 4. *Tokens (Words) Extracted from Incidents Dataset*

```
> colnames (iw3)
  [1] "type"           "fire"          "gunman"      "killed"       "opened"      "police"        "apartment"     "children"
  [9] "died"           "fatally"       "girlfriend"  "inside"       "later"       "shot"          "three"         "man"
 [17] "night"          "one"           "two"         "women"        "friends"     "home"          "life"          "took"
 [25] "wife"           "according"     "arrested"    "authorities"  "charged"     "connection"    "days"          "family"
 [33] "five"           "house"         "killing"     "murders"      "neighbors"   "rifle"         "several"       "shooting"
 [41] "went"           "allegedly"     "fired"       "injuring"     "parents"     "people"        "residence"     "four"
 [49] "others"         "party"         "victims"     "assailant"    "injured"     "adults"        "woman"         "back"
 [57] "death"          "entered"       "outside"     "sons"         "committing"  "day"           "also"          "child"
 [65] "dead"           "dispute"       "found"       "reportedly"   "responding"  "men"           "murdersuicide" "another"
 [73] "call"           "counts"        "discovered"  "father"       "murder"      "called"        "members"       "drove"
 [81] "related"        "seven"         "shootings"   "standoff"     "believe"     "six"           "sister"        "survived"
 [89] "eight"          "domestic"      "history"     "mother"       "violence"    "suicide"       "lee"           "bodies"
 [97] "suspect"        "stabbed"       "wounded"     "case"         "charges"     "exgirlfriend"  "order"         "time"
[105] "investigators"  "county"        "incident"    "including"    "officer"     "drug"          "james"         "remains"
[113] "fourth"         "hospital"      "robbery"     "say"          "scene"       "vehicle"       "left"          "awaiting"
[121] "trial"          "handgun"       "set"         "daughter"     "daughters"   "returned"      "dropped"       "injuries"
[129] "officers"       "michael"       "guilty"      "parole"       "pleaded"     "sentence"      "without"       "committed"
[137] "mental"         "said"          "young"       "believed"     "unsolved"    "ended"         "convicted"     "prison"
[145] "sentenced"      "eligibility"   "ages"        "former"       "son"         "received"      "brother"       "used"
[153] "prior"          "months"        "car"         "years"        "given"       "earlier"       "estranged"     "kids"
[161] "told"           "relatives"     "sentences"   "kill"         "boyfriend"   "couple"        "serving"       "slayings"
```

EXPLORATION OF MACHINE LEARNING METHODS USING TF-IDF WEIGHTED TOKENS

Figure 5. *Factor Levels for the Multi-Class Outcome*

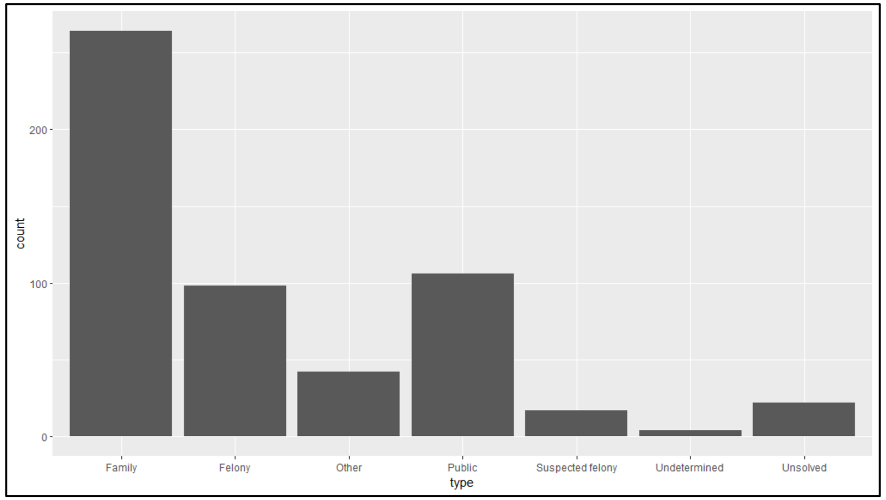


Figure 6. *Accuracy and Kappa Statistics by Model*

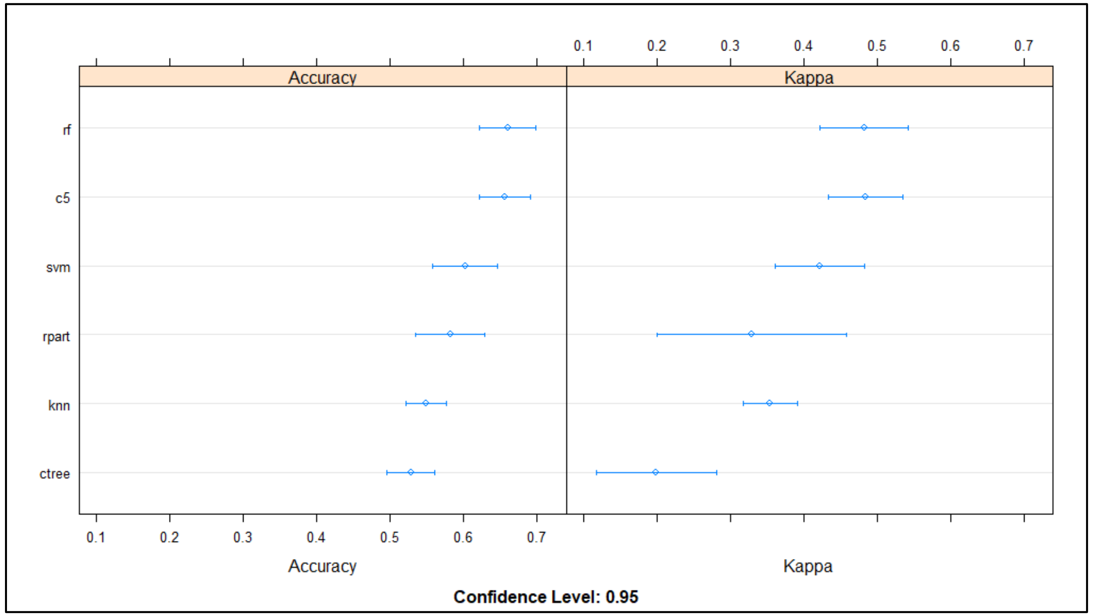
```
Call:
summary.resamples(object = results)

Models: ctree, rpart, c5, svm, knn, rf
Number of resamples: 10

Accuracy
           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
ctree 0.4629630 0.4888001 0.5456349 0.5286604 0.5599415 0.5849057    0
rpart 0.4814815 0.5507519 0.5857700 0.5821597 0.6284722 0.6792453    0
c5    0.5555556 0.6315789 0.6666667 0.6565030 0.6755952 0.7358491    0
svm   0.5178571 0.5555556 0.5862573 0.6027475 0.6578164 0.6851852    0
knn   0.4814815 0.5231481 0.5584795 0.5493028 0.5815364 0.5964912    0
rf    0.5925926 0.6266447 0.6522989 0.6607141 0.7083333 0.7358491    0

Kappa
            Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
ctree 0.01136364 0.1161165 0.2097977 0.1988840 0.2706449 0.4065934    0
rpart 0.00000000 0.3345313 0.3906120 0.3288896 0.4277723 0.5142857    0
c5    0.32919255 0.4608912 0.4786608 0.4842673 0.5197922 0.5927552    0
svm   0.30990415 0.3466647 0.4063818 0.4218467 0.5071457 0.5248447    0
knn   0.26694717 0.3199713 0.3735661 0.3542111 0.3809188 0.4307425    0
rf    0.33926585 0.4495966 0.4680049 0.4824585 0.5490586 0.5998094    0
```

Figure 7. *Accuracy and Kappa Statistics and 95%CIs by Model*



RESULTS: RANDOM FOREST

```
> set.seed (1)
> start.time <- Sys.time()
> fit.rf <- train(type~., data=iw3, method="rf", metric=metric, trControl=control)
> fit.rf
Random Forest

553 samples
167 predictors
  7 classes: 'Family', 'Felony', 'Other', 'Public', 'Suspected felony', 'Undetermine

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 499, 499, 499, 496, 496, 500, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
    2   0.6259195  0.3641152
   84   0.6588623  0.4788893
  167   0.6424950  0.4548221

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 84.
```

Figure 8. *Random Forest Modeling Output*

Figure 9. *Updated Accuracy and Kappa Metrics through Use of Tunegrid (Random Forest)*

```
Random Forest

553 samples
167 predictors
  7 classes: 'Family', 'Felony', 'Other', 'Public', 'Suspected felony', 'Undetermined', 'Unsolved'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 498, 498, 497, 498, 498, 496, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
    2   0.6237217  0.3626466
    7   0.6783682  0.4902298
   12   0.6928487  0.5175811
   17   0.6835617  0.5074648
   22   0.6765441  0.4984083
   27   0.6765789  0.4997332
   32   0.6748570  0.4984393
   37   0.6710582  0.4927903
   42   0.6783008  0.5032329
   47   0.6820323  0.5121947
   52   0.6783321  0.5060217
   57   0.6854448  0.5185433
   62   0.6873291  0.5211522
   67   0.6801190  0.5102410
   72   0.6692412  0.4941472
   77   0.6693387  0.4943471
   82   0.6729414  0.5012998
   87   0.6692088  0.4960176
   92   0.6746018  0.5037500
   97   0.6620660  0.4863461
  102   0.6692749  0.4952054
  107   0.6710293  0.4966127
  112   0.6818746  0.5143193
  117   0.6602802  0.4829721
  122   0.6746320  0.5059979
  127   0.6747294  0.5059665
  132   0.6674254  0.4920492
  137   0.6638191  0.4867558
  142   0.6729739  0.5024602
  147   0.6673929  0.4926138
  152   0.6656060  0.4877206
  157   0.6727813  0.5009800
  162   0.6548268  0.4738816
  167   0.6673581  0.4916982

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 12.
```

RESULTS: RANDOM FOREST

Figure 10. *Accuracy and Kappa Metrics at Varying mtry*

```
Random Forest

553 samples
167 predictors
  7 classes: 'Family', 'Felony', 'Other', 'Public', 'Suspected felony', 'Undetermined', 'Unsolved'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 499, 499, 499, 496, 496, 500, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
    2   0.6097562  0.3323045
   12   0.6785201  0.4913957
   22   0.6732872  0.4914016
   32   0.6660701  0.4828711
   42   0.6625650  0.4803571
   52   0.6678291  0.4885215
   62   0.6641602  0.4833982
   72   0.6585396  0.4757806
   82   0.6625660  0.4830511
   92   0.6499384  0.4652658
  102   0.6625383  0.4817969
  112   0.6444454  0.4561968
  122   0.6534426  0.4684575
  132   0.6464599  0.4599118
  142   0.6496692  0.4633105
  152   0.6532486  0.4718375
  162   0.6460675  0.4592269

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 12.
```

Figure 11. *Feature Importance by Token: Top 20*

```
> # Calculate feature importance
> importance <- varImp(fit.rf)
> # Print the importance data
> print(importance)
rf variable importance

  only 20 most important variables shown (out of 167)

          Overall
children   100.00
wife        72.35
four        55.44
family      55.08
men         48.89
people      44.77
found       41.83
home        39.59
case        38.37
shot        37.97
drug        37.32
killed      36.77
life        36.76
two         34.33
sentenced   34.29
robbery     33.95
police      32.54
death       31.47
three       30.74
others      30.48
```

RESULTS: SVM

Figure 12. *Accuracy and Kappa Metrics at Varying Cost*

```
> set.seed(111)
> start <- proc.time()
> train_control <- trainControl(method="cv",number=10)
> svm_grid <- expand.grid(C=seq(0.1,3.1, length=5))
> fit.svm <- train(type~., data=iw3, method="svmLinear", trControl=train_control, tuneGrid=svm_grid)
> fit.svm
Support Vector Machines with Linear Kernel

553 samples
167 predictors
  4 classes: 'Family', 'Felony', 'Public', 'Other'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 497, 498, 496, 497, 499, 499, ...
Resampling results across tuning parameters:

  C     Accuracy   Kappa
  0.10  0.6404004  0.4646951
  0.85  0.6276648  0.4430787
  1.60  0.6276648  0.4430787
  2.35  0.6276648  0.4430787
  3.10  0.6276648  0.4430787

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was C = 0.1.
```

Figure 13. Accuracy by Cost

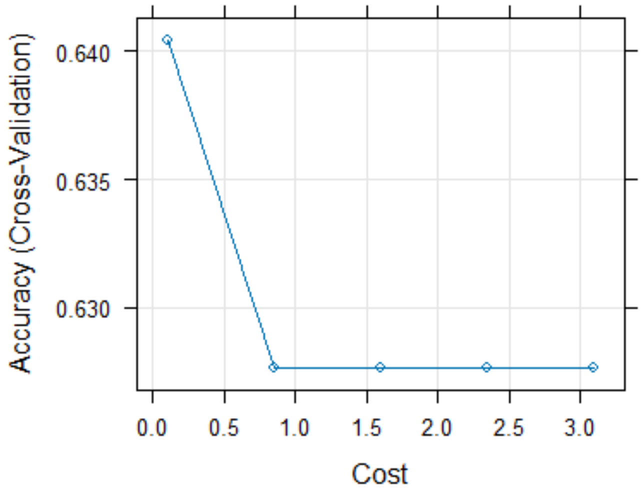
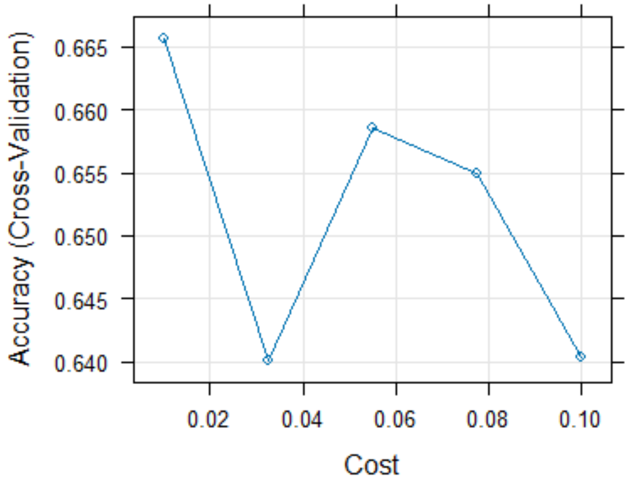


Figure 14. Accuracy by Cost – Soft Margin



RESULTS: C5

Figure 15. *C5 Model Parameters and Performance*

```
> set.seed (1)
> start.time <- Sys.time()
> fit.c5 <- train(type~., data=iw3, method="C5.0", metric=metric, trControl=control)

> fit.c5
C5.0

553 samples
167 predictors
  4 classes: 'Family', 'Felony', 'Public', 'Other'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 498, 498, 499, 498, 497, 497, ...
Resampling results across tuning parameters:

  model  winnow  trials  Accuracy   Kappa
  rules  FALSE    1      0.5731906  0.3361197
  rules  FALSE   10      0.6692124  0.4990592
  rules  FALSE   20      0.6617819  0.4908866
  rules   TRUE    1      0.5605968  0.3222386
  rules   TRUE   10      0.6061163  0.3918448
  rules   TRUE   20      0.6005342  0.3814689
  tree   FALSE    1      0.5587715  0.3447382
  tree   FALSE   10      0.6293422  0.4407611
  tree   FALSE   20      0.6401574  0.4587260
  tree    TRUE    1      0.5751735  0.3475118
  tree    TRUE   10      0.5894615  0.3669209
  tree    TRUE   20      0.5822213  0.3584054

Accuracy was used to select the optimal model using the largest value.
The final values used for the model were trials = 10, model = rules and winnow = FALSE.
```

Figure 16. *Parameter Tuning Profile over Boosted Iterations*

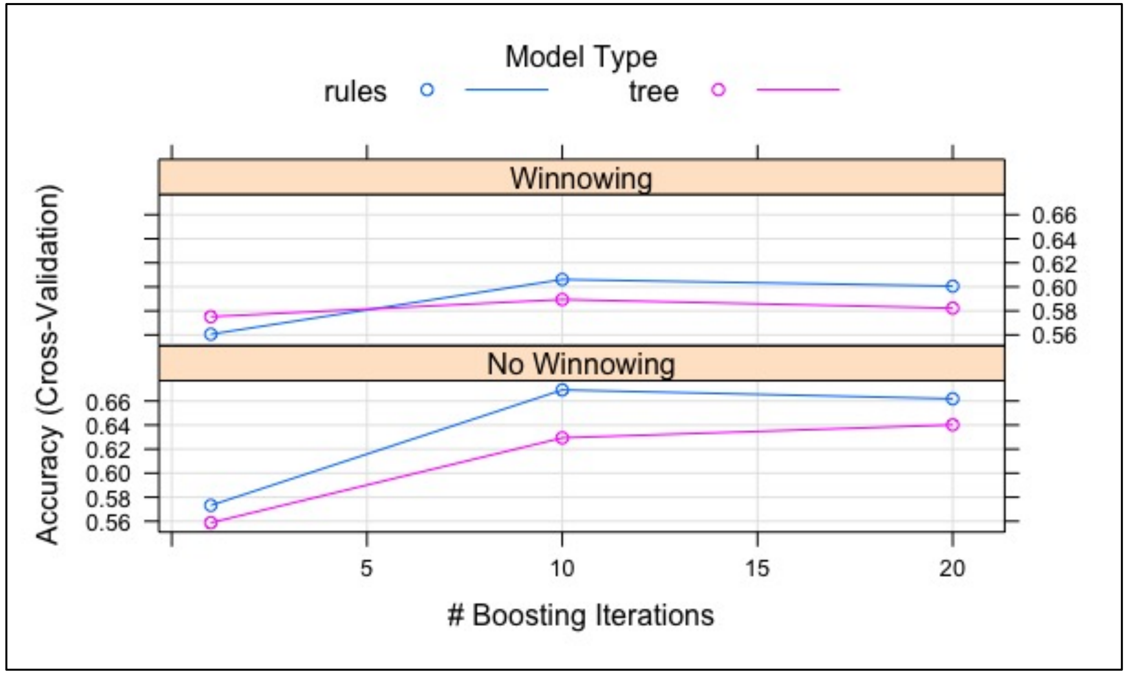


Figure 17. Test Classification Results

```
 (a)   (b)   (c)   (d)    <-classified as
 ----  ----  ----  ----
 264                      (a): class Family
        98                (b): class Felony
   1         105          (c): class Public
   2                83    (d): class Other
```

RESULTS: C5

Figure 18. *Attributes, Usage Rates and Rulesets*

```
Rule 9/10: (10.9/1, lift 2.4)
        injuring <= 0.1068867
        party <= 0.1157907
        eligibility > 0.106482
        relatives <= 0
        ->  class Family  [0.842]

Rule 9/11: (7.2/0.5, lift 2.4)
        injuring <= 0.1068867
        sister > 0
        ->  class Family  [0.841]

Rule 9/12: (9/0.8, lift 2.4)
        day <= 0
        history <= 0
        prior > 0
        ->  class Family  [0.836]

Rule 9/13: (12.3/1.4, lift 2.4)
        drove <= 0.1575242
        case <= 0.1246196
        relatives > 0
        ->  class Family  [0.832]

Rule 9/14: (10.8/1.2, lift 2.4)
        call > 0.05847466
        ->  class Family  [0.831]

Rule 9/15: (10.3/1.3, lift 2.3)
        house <= 0.1394604
        sons > 0.07611626
        bodies <= 0.03803859
        serving <= 0
        ->  class Family  [0.811]

Rule 9/16: (14.2/2.3, lift 2.3)
        party <= 0.1157907
        also > 0.1187697
        case <= 0.1246196
        robbery <= 0
        son <= 0.07558072
        ->  class Family  [0.797]
```

Figure 19. *Word Inclusions in C5*

```
Attribute usage:

100.00% wife
100.00% case
100.00% unsolved
 99.64% children
 98.01% dead
 90.42% years
 90.24% family
 85.71% robbery
 84.09% drug
 83.54% daughter
 75.95% house
 75.59% history
 75.23% men
 65.46% life
 65.46% vehicle
 63.83% mother
 61.12% counts
 61.12% charges
 60.58% injuring
 60.04% parents
 59.86% girlfriend
 58.05% opened
 56.78% two
 56.24% father
 56.06% dropped
 54.25% returned
 54.07% remains
 54.07% boyfriend
 53.35% injured
 52.80% took
 52.62% bodies
 52.26% entered
 52.26% drove
 52.08% death
 50.99% allegedly
 50.45% exgirlfriend
 49.91% county
 49.01% according
 48.28% murders
 46.65% five
 46.65% seven
 46.29% members
 46.11% killing
 45.93% another
 45.57% gunman
 45.03% prison
 44.85% time
 44.85% pleaded
 44.30% survived
 43.04% party
 43.04% relatives
 42.13% sons
 42.13% received
 41.05% serving
 40.33% authorities
 39.78% residence
 39.60% investigators
 39.24% believe
 39.24% awaiting
 38.16% rifle
 38.16% ended
 37.97% lee
 37.79% victims
 37.79% son
 37.61% discovered
 37.25% several
 37.25% convicted
 37.07% found
 36.89% outside
 36.53% sentenced
 35.99% daughters
 35.99% months
 34.90% back
 34.54% also
 34.54% handgun
 34.36% responding
 34.36% sister
 34.00% officers
 34.00% kids
 33.82% left
 33.45% one
 32.91% murder
 31.46% shootings
 31.28% went
 30.92% believed
 30.20% former
 29.29% man
 29.29% child
 28.93% arrested
 28.75% friends
 28.39% set
 28.21% woman
 27.49% related
 26.40% eight
 26.04% home
 25.68% scene
 24.95% stabbed
 24.77% connection
 24.77% reportedly
 24.59% later
 24.23% killed
 24.05% shot
 24.05% prior
 23.87% fourth
 23.51% car
 23.33% charged
 22.97% neighbors
 22.97% called
 22.97% couple
 22.78% young
 22.42% murdersuicide
 22.42% suspect
 22.24% wounded
 21.34% others
 20.98% call
 20.80% women
 20.61% eligibility
 19.89% say
 19.35% michael
 18.63% domestic
 18.44% fire
 17.72% apartment
 17.18% days
 17.00% said
 16.64% fired
 16.64% earlier
 15.91% hospital
 14.83% james
 14.10% trial
 14.10% injuries
 14.10% used
 13.02% dispute
 12.84% including
 12.12% sentence
 11.75% brother
 11.39% died
 11.39% slayings
 10.49% suicide
 10.13% inside
  9.76% order
  8.68% adults
  8.68% incident
  8.68% mental
  7.41% four
  6.15% told
  5.24% ages
  4.16% assailant
  4.16% guilty
  3.80% three
  3.44% police
  3.25% committing
  2.89% fatally
  2.35% shooting
  2.17% day
  1.81% committed
  1.63% six
  0.36% people
```

RESULTS: MULTI-MODEL COMPARISON (1)

Figure 20. *Factor Levels for the Revised Multi-Class Outcome*

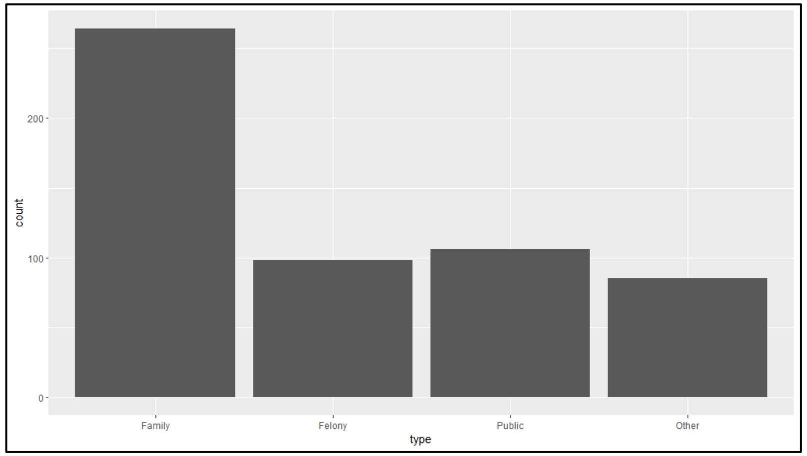


Figure 21. *Accuracy and Kappa Statistics by Model*

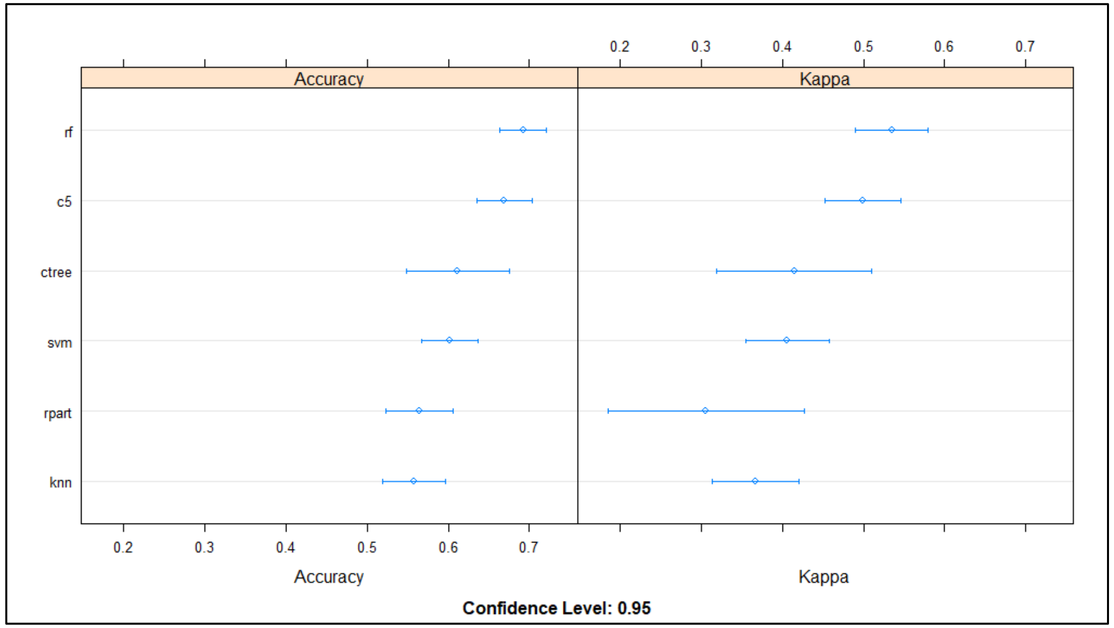
```
Call:
summary.resamples(object = results)

Models: ctree, rpart, c5, svm, knn, rf
Number of resamples: 10

Accuracy
           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
ctree 0.4912281 0.5381494 0.5995370 0.6116926 0.6863636 0.7454545    0
rpart 0.4727273 0.5381494 0.5688552 0.5640695 0.5928230 0.6727273    0
c5    0.5818182 0.6409091 0.6726190 0.6692124 0.7005051 0.7454545    0
svm   0.4909091 0.5836851 0.6126623 0.6023187 0.6267677 0.6727273    0
knn   0.4909091 0.5157468 0.5503367 0.5583574 0.5792208 0.6666667    0
rf    0.6363636 0.6681818 0.6901629 0.6926213 0.7227273 0.7636364    0

Kappa
           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
ctree 0.1865157 0.3186726 0.3967680 0.4144725 0.5163147 0.6122860    0
rpart 0.0000000 0.3185318 0.3622589 0.3053897 0.3935753 0.5032614    0
c5    0.3918269 0.4480633 0.5096433 0.4990592 0.5445779 0.6045198    0
svm   0.2380010 0.3864243 0.4143520 0.4059962 0.4293327 0.5151812    0
knn   0.2797007 0.3170686 0.3433910 0.3663960 0.3895107 0.5095109    0
rf    0.4249869 0.4968788 0.5435568 0.5350023 0.5700759 0.6372400    0
```

Figure 22. *Accuracy and Kappa Statistics and 95%CIs by Model*



RESULTS: MULTI-MODEL COMPARISON (2)

Figure 23. *Pair-wise Model Comparisons, Bonferroni Adjusted*

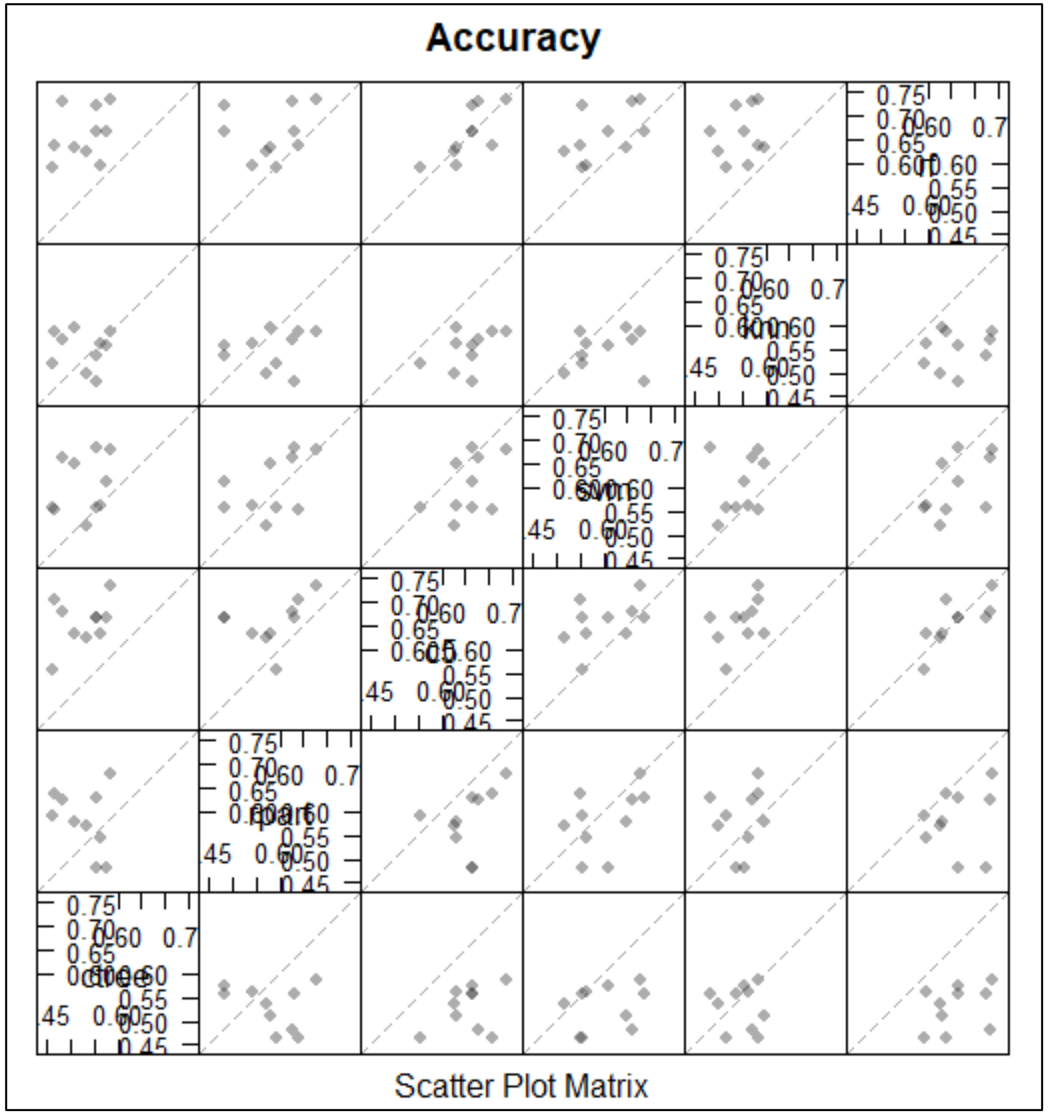
```
Call:
summary.diff.resamples(object = diffresults)

p-value adjustment: bonferroni
Upper diagonal: estimates of the difference
Lower diagonal: p-value for H0: difference = 0

Accuracy
      ctree     rpart     c5        svm       knn       rf
ctree           -0.053499 -0.127843 -0.074087 -0.020642 -0.132054
rpart 1.0000000           -0.074343 -0.020588  0.032857 -0.078554
c5    0.0005424 0.0982717            0.053756  0.107200 -0.004211
svm   0.0992756 1.0000000 0.2693491            0.053445 -0.057967
knn   1.0000000 1.0000000 0.0008483 0.4518839           -0.111411
rf    0.0007762 0.1684529 1.0000000 0.1379767 0.0032738

Kappa
      ctree     rpart     c5        svm       knn       rf
ctree           -0.130006 -0.285383 -0.222963 -0.155327 -0.283575
rpart 1.0000000           -0.155378 -0.092957 -0.025321 -0.153569
c5    0.0004294 0.3338188            0.062421  0.130056  0.001809
svm   0.0048736 1.0000000 0.5662214            0.067636 -0.060612
knn   0.0182199 1.0000000 0.0033684 0.5232951           -0.128247
rf    0.0005827 0.6695741 1.0000000 0.6786959 0.0156914
```

Figure 24. *Scatterplot – Accuracy and Select Machine Learning Methods*



CONCLUSIONS

Observation 1: The use of text mining methods in conjunctions with machine learning models presents a novel and powerful way to connect narratives to outcomes associated with mass killings – as a central component of this analysis, the creation of an integrated data frame with tokenized and weighted components was a logical and straightforward process also consistent with the text mining literature and the idea that non-structured, text data is increasingly important in data mining and machine learning.

Observation 2: Data wrangling with respect to the structuring and balancing of the outcome variable makes an immediate difference in resampling accuracy; pair-wise comparisons of model performance using statistical methods can provide insights that are not immediately apparent with the generation of individual model summaries. Although use of resampling accuracy is a logical first step in this type of analysis, generation of subsets (e.g., train, test) and further refinement of the sampling process to adjust for imbalances would be ideal in the continuation of this work.

Observation 3: From a learning perspective, the inclusion of random forests and other algorithms in conjunction with the text mining process in a critically important social context – and through the team's use of iterative and incremental cycles that generated this work – brought together multiple aspects of data mining that this course introduced.