Predicting threatening situations in custody-related child abductions

The Protectors: May Chan, Padma Hari, Kimberly Sanders, Marjorie Waters

Friday, September 19, 2014

Our ultimate goal statement is to identify the risk factors that lead to custodial/non-custodial child abduction. The data available for this crime is extremely limited and at this time we have only a small sample of 20 incidents from 2013. We are expecting an additional 3-4 years of data later in September. Based on conversations with our FBI sponsor, our secondary goal is to determine the factors that indicate higher risk of a threatening situation developing and that will be the focus of our initial exploration.

The preliminary data set was provided by the FBI. In addition to standard options, the string "U" for unidentified was also converted to NA for consistent treatment in R.

For the purposes of this in class, all R-code is shown along with the output, even for the plots. Messages and output have been turned off for functions with substantial commentary such as geocoding from ggmaps and running the models for performance estimation. For a more polished version to share with sponsors, echo will be turned to FALSE for much of the code..

```r
require (DMwR, quietly=T)
require (ggplot2, quietly=T)
require (e1071, quietly=T)
require (performanceEstimation, quietly=T)
require (randomForest, quietly=T)
require (rpart, quietly=T)
require (plyr, quietly=T)
require (corrplot,quietly=T)
require (ggmap,quietly=T)
require (gridExtra,quietly=T)
```

```r
# Load the data into R and convert column names
data.raw <- read.csv("Sample_data.csv",
                     stringsAsFactors=TRUE, na.strings = c("NA","U",""," "))
names.data <- c("Date","Victim.Age","Victim.Race","Victim.Gender","Harm",
                "Publicized","Location","Region","Relationship","RSO","Offender.Age",
                "Offender.Race","Offender.Gender","Rural.City","Missing.Location",
                "Recovery.Location", "Recovery", "Number.Victims")
colnames(data.raw) <- names.data

# Remove all carriage returns
for (i in 1:length(data.raw))
{
  if (class(data.raw[,i]) == "factor")
    {levels(data.raw[,i]) <- gsub("\n","", levels(data.raw[,i]))}
}

head(data.raw,1)
```

```
##     Date Victim.Age Victim.Race Victim.Gender Harm Publicized   Location
## 1 13-Aug          8       Black        Female   No        Yes Birmingham
##          Region       Relationship RSO Offender.Age Offender.Race
## 1 South Central Father'sGirlfriend  No           35         Black
##   Offender.Gender Rural.City Missing.Location
## 1          Female      Rural             <NA>
##               Recovery.Location Recovery Number.Victims
## 1 >10 miles fromMissingLocation    Alive              1
```

Data pre-processing

Missing data

The data set was evaluated for missing data. Several attributes are considered critical to the analysis, victim attributes and relationship of victim to offender; observations with insufficient basic information were removed.

```
# Remove lines where victim info and relationship is not provided, one line removed
data <- data.raw[complete.cases(data.raw[,c(2,3,4,5,9)]),]
```

This results in the removal of 1 line(s). For the other attributes, the number of missing fields was determined to help determine the best approach for handling them.

```
# Count NA by variable and isolate fields with at least 1 NA
data.empty = NULL
for (i in 1:length(data))
{
  data.empty[i] <- length(which(is.na(data[i])))
}
x <- as.data.frame(cbind(names.data, data.empty)[order(-data.empty),])
x <- x[which(x$data.empty != 0),]
```

names.data	data.empty
Recovery.Location	10
Missing.Location	9
RSO	5
Offender.Race	2
Offender.Age	1
Rural.City	1

Factors with ~50% or more missing instances, Recovery.Location and Missing.Location, were removed from the data set. The other missing data was filled using k-nearest neighbor matching, using k=3.

```
# Remove Recovery.Location and Missing.Location, almost half of observations missing data
data <- data[,-which(names(data) %in% c("Recovery.Location","Missing.Location"))]

# Fill RSO, Offender.Race, Offender.Age, Rural.City using nearest neighbors
data <- knnImputation(data, k=3)
check <- nrow(data[!complete.cases(data),])
```

Generalized groupings

Age groupings of the victim and offender were created. This aids in data evaluations as we are interested in differences between younger children and older children. In addition, our FBI sponsor is interested in generalized risk factors.

```
# Create victim age group
data$Victim.AgeGroup <- cut(data$Victim.Age,
                            breaks=c(-0.5,5.5,10.5,Inf),
                            labels=c('0-5','6-10','11-15'))
table(data$Victim.AgeGroup)
```

```
##
##   0-5  6-10 11-15
##    10     6     5
```

```
data$Offender.AgeGroup <- cut(data$Offender.Age,
                            breaks=c(20,30,40,50,Inf),
                            labels=c('<30','30-40','40-50','>50'))

table(data$Offender.AgeGroup)
```

```
##
##   <30 30-40 40-50   >50
##     5    10     5     1
```

A relationship group was created to generalize the more detailed list provided in the original data. As an example, “Father’s girlfriend” and “Mother’s boyfriend” were both converted to the group “Family friend”.

```
relate.list <- as.data.frame(unique(unlist(data$Relationship), use.names=FALSE))
colnames(relate.list) <- c("Relationship")
relate.list$Relate.Group <- as.factor(c("Family Friend", "Relative", "Parent","Parent",
      "Family Friend","Family Friend","Family Friend","Family Friend","Relative","Parent"))
data <- merge(data, relate.list, by = "Relationship")
table(data$Relate.Group)
```

```
##
## Family Friend        Parent      Relative
##             5            14             2
```

Creating the target variable

The analysis is intended to identify abductions with a high risk of potential violence. While actions from law enforcement most often prevent tragic outcomes, these situations may escalate to physical or psychological harm to the child. Identifying these situations in advance would be beneficial for focusing efforts. Our target variable is thus situations with a threat of violence vs those without. For example, incidents where the child was abducted at gunpoint is considered violent, even if the child was not physically harmed. Our sample has no instances where the child was physically harmed, in the larger data this would be included.

```
# Create a list of harm responses and identify them as 1/0
harm.list <- as.data.frame(unique(unlist(data$Harm), use.names=FALSE))
colnames(harm.list) <- c("Harm")
harm.list$target <- c(0,1,1,1,0,1)
data <- merge(data, harm.list, by = "Harm")
table(data$target)
```

```
##
##  0  1
## 15  6
```

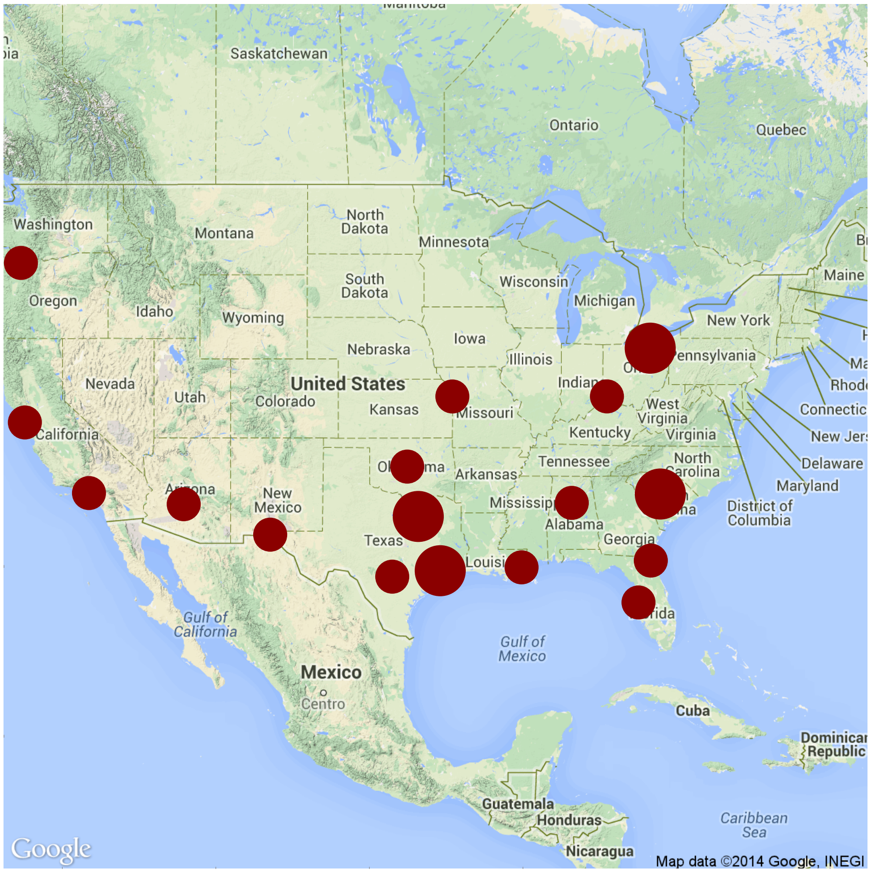
For the small sample data set, 6 out of 21 incidents (28.6%) show evidence of threatening situations.

Evaluating the data

Geographic distribution

We looked at the geographic distribution of incidents in the data, the size of the dot indicates the number of incidents in that city. The incidents are generally spread across the US, although there are no examples from the northeast or the northwest states. The majority of incidents occur in the south and central regions.

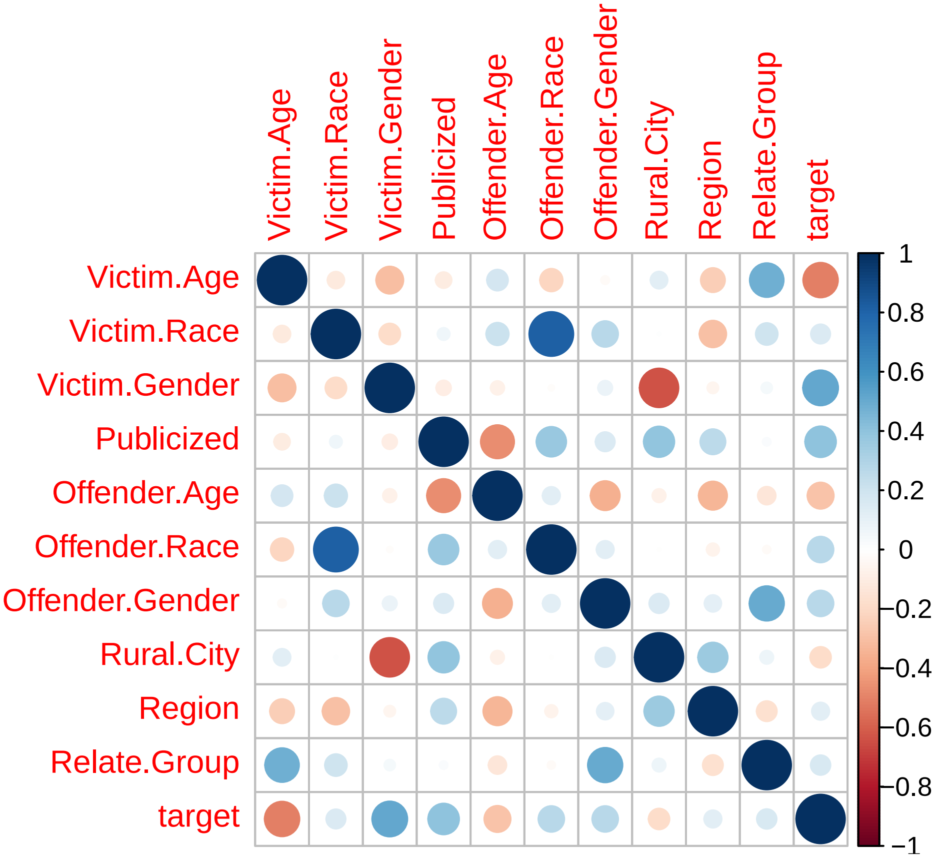
```
library(ggmap)
loc <- data.frame()
all_locs <- paste0(as.character(data$Location),", US")
unique_locs <- unique(all_locs)
for (i in 1:length(unique_locs))
{
  geo.res <- geocode(unique_locs[i], messaging = FALSE)
  loc[i,1] <- geo.res[1]
  loc[i,2] <- geo.res[2]
  loc[i,3] <- length(all_locs[all_locs==unique_locs[i]])
}

map <- get_map("united states", zoom = 4)
ggmap(map, extent="device") + geom_point(aes(x=lon, y=lat),
              data=loc, color="darkred", size=3+3*loc$V3)
```



Correlations

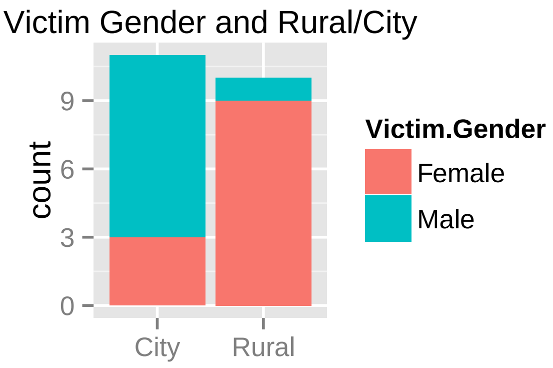
The correlations of several key attributes are shown below.

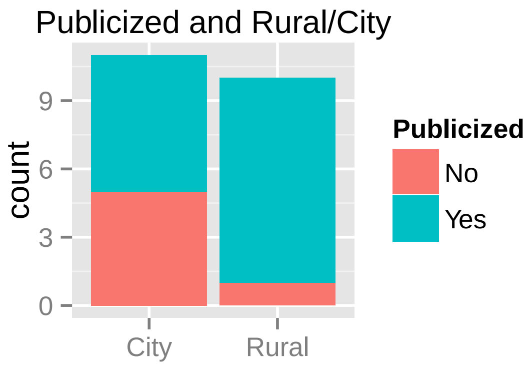
```
cor.data <- data[,c(4:7,11:14,9,19,20)]
for (i in 1:length(cor.data))
{
  cor.data[,i] <- as.numeric(cor.data[,i])
}
c <- cor(cor.data)
corrplot(c)
```

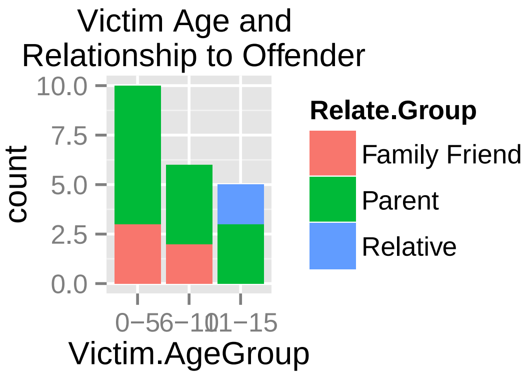


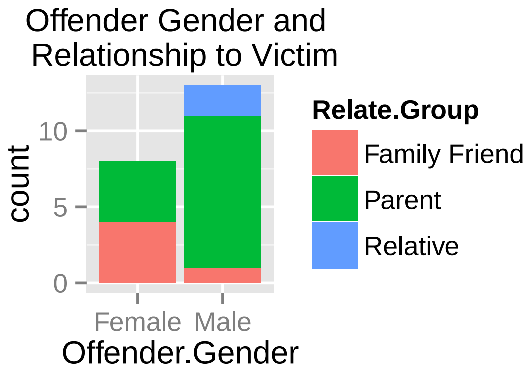
Interestingly from a data perspective, rural.city is correlated to victim gender and publicized; rural incidents in our sample data set are more likely to be publicized, incidents in cities are more often male victims. Victim age and offender gender are slightly correlated to victim-offender relationship.

```
# detailed bar charts of interesting correlations
p1 <- ggplot(data, aes(x=Rural.City, fill=Victim.Gender)) + geom_bar(stat='bin') +
  ggtitle("Victim Gender and Rural/City") +
  theme(plot.title=element_text(size=12), axis.title.x=element_blank())
p2 <- ggplot(data, aes(x=Rural.City, fill=Publicized)) + geom_bar(stat='bin') +
  ggtitle("Publicized and Rural/City") +
  theme(plot.title=element_text(size=12), axis.title.x=element_blank())
p3 <- ggplot(data, aes(x=Victim.AgeGroup, fill=Relate.Group)) + geom_bar(stat='bin') +
  ggtitle("Victim Age and \n Relationship to Offender") +
  theme(plot.title=element_text(size=12))
p4 <- ggplot(data, aes(x=Offender.Gender, fill=Relate.Group)) + geom_bar(stat='bin') +
  ggtitle("Offender Gender and \n Relationship to Victim") +
  theme(plot.title=element_text(size=12))
grid.arrange(p1,p2,p3,p4,ncol=2)
```



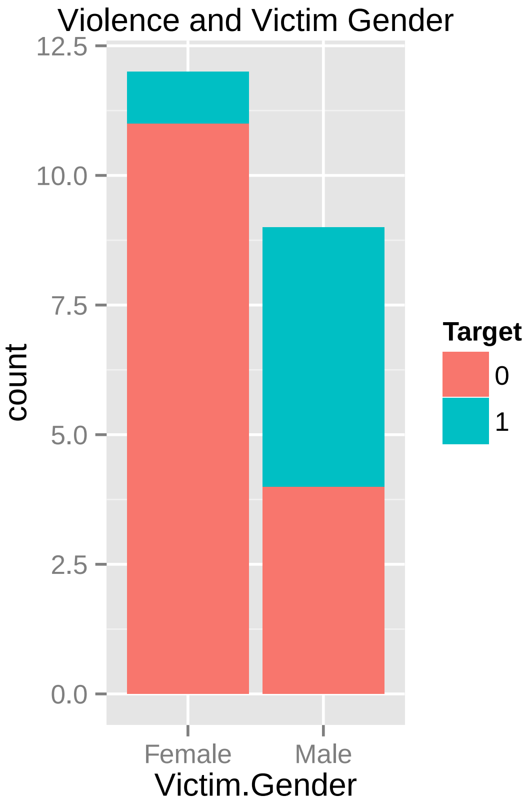


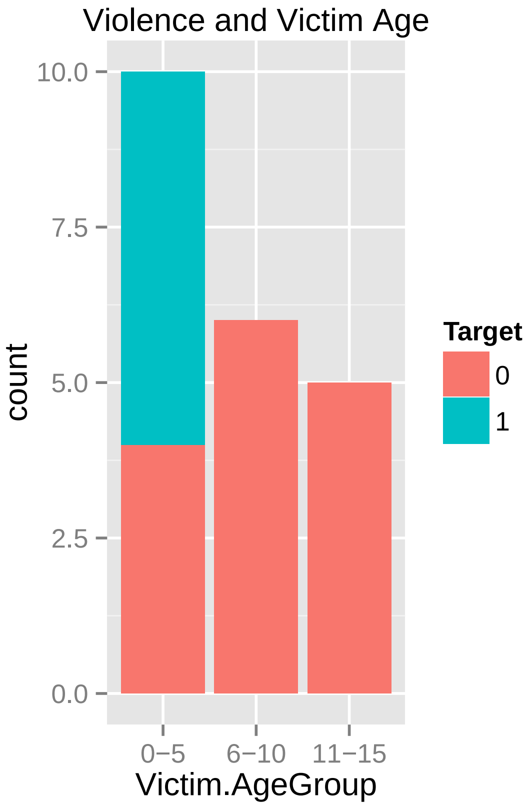




Interestingly, Victim age and Victim gender are correlated to the target. Although there are more females victims represented in the sample, males are more likely to be exposed to incidents with violence. All of the violent incidents are with children 5 and under.

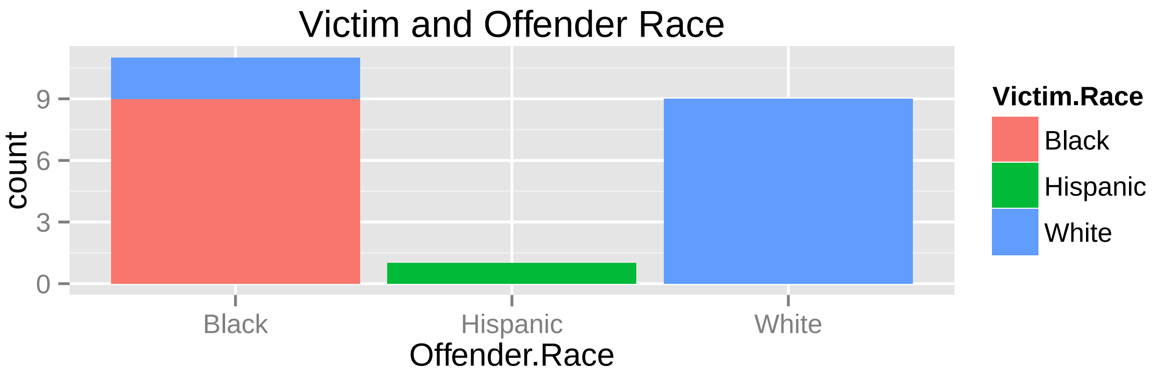
```
# detailed bar charts of interesting correlations witht he target variable
p1 <- ggplot(data, aes(x=Victim.Gender, fill=as.factor(target))) + geom_bar(stat='bin') +
  ggtitle("Violence and Victim Gender") +
  theme(plot.title=element_text(size=12)) + scale_fill_discrete(name="Target")
p2 <- ggplot(data, aes(x=Victim.AgeGroup, fill=as.factor(target))) + geom_bar(stat='bin') +
  ggtitle("Violence and Victim Age") +
  theme(plot.title=element_text(size=12)) + scale_fill_discrete(name="Target")
grid.arrange(p1,p2, ncol=2)
```

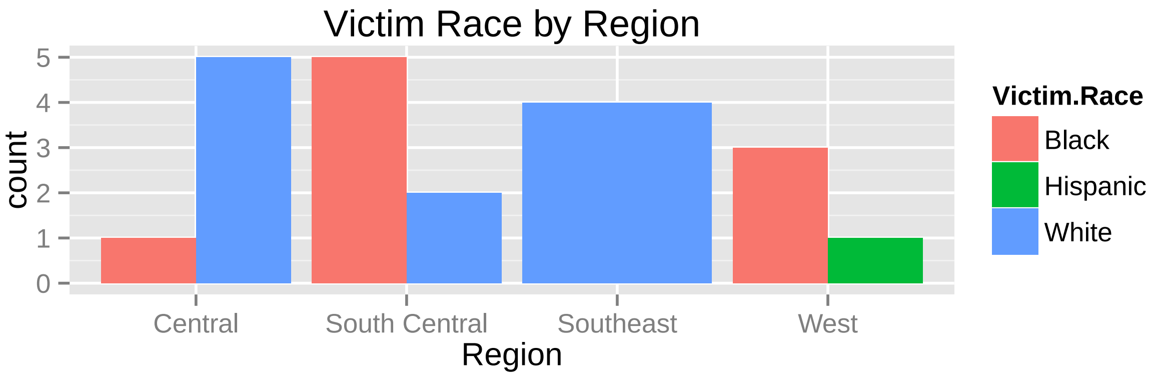




The race of the offender and victim is highly correlated, which is unsurprising. Looking further into the data, White and Black offenders dominate abduction cases with whites dominating in Southeast, Central and blacks dominating in South Central and West, Hispanic abduction appears rare in our data from our additional research with DCFS this is may be due to non reporting and immigration status. Our FBI sponsor has cautioned about including race and region in the analysis as this are politically sensitive topics, however from a data driven perspective we have maintained them for the analysis.

```
p1 <- ggplot(data, aes(x=Offender.Race, fill=Victim.Race)) + geom_bar(stat='bin') +
  ggtitle("Victim and Offender Race")
p2 <- ggplot(data, aes(x=Region, fill=Victim.Race)) + geom_histogram(binwidth=.5, position="dodge") +
  ggtitle("Victim Race by Region")
grid.arrange(p1,p2, ncol=1)
```

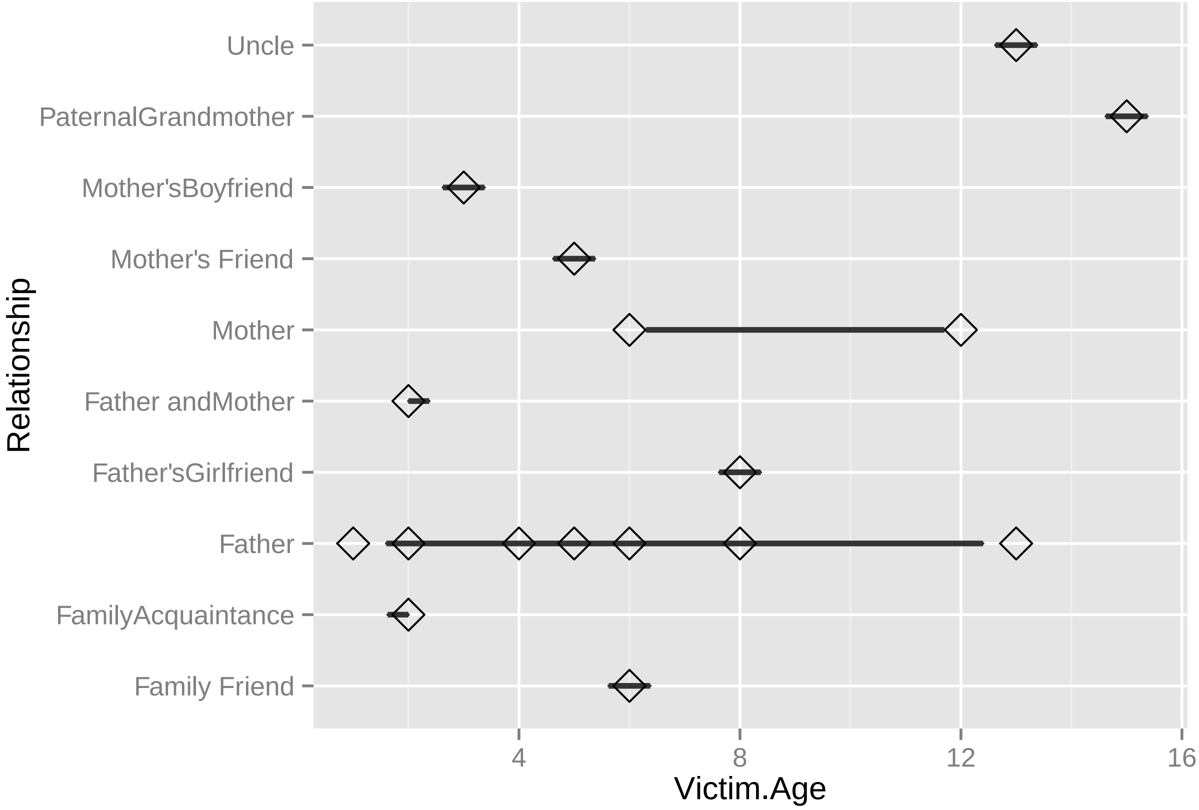




Additional visual data understanding

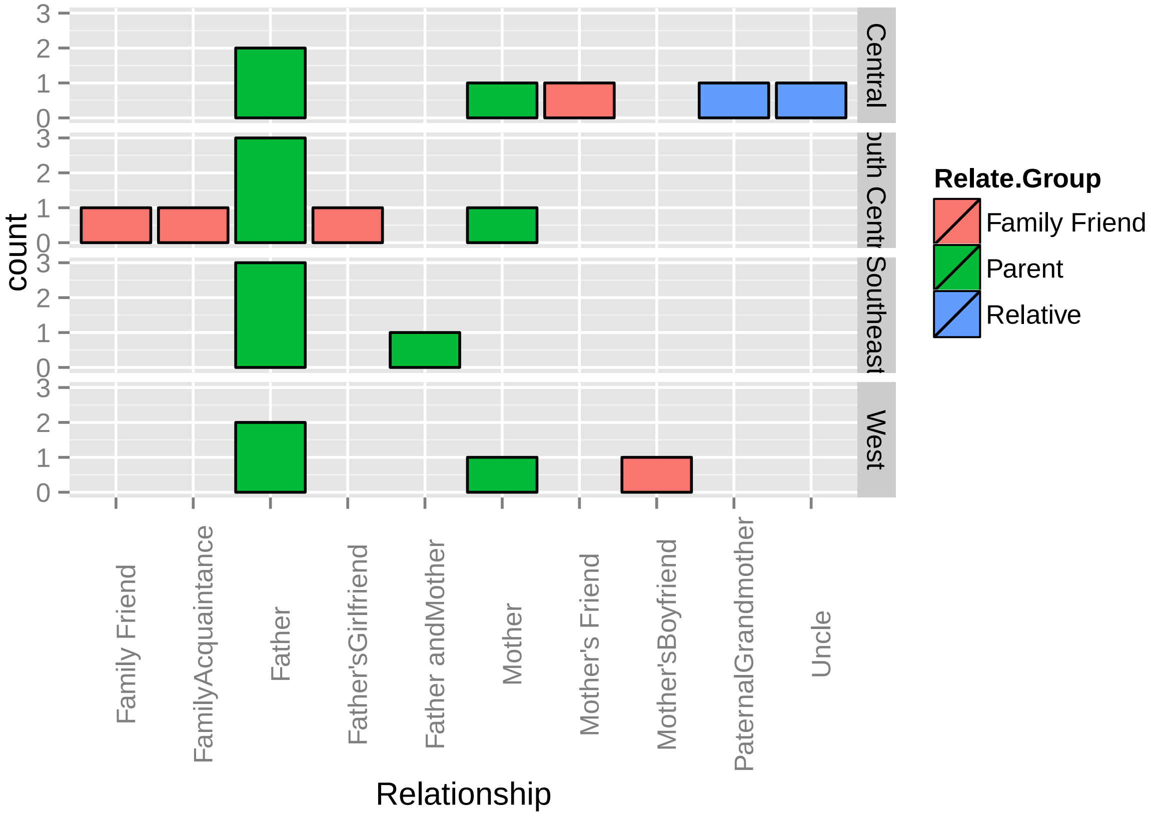
The father appears to have the highest incident rate in our limited sample set, however, statically speaking, based on additional information from DCFS, the mother is more often the offender.

```
ggplot(data, aes(x=Victim.Age, y=Relationship)) + geom_boxplot() +
  stat_summary(fun.y=mean, geom="point", shape=5, size=4)
```



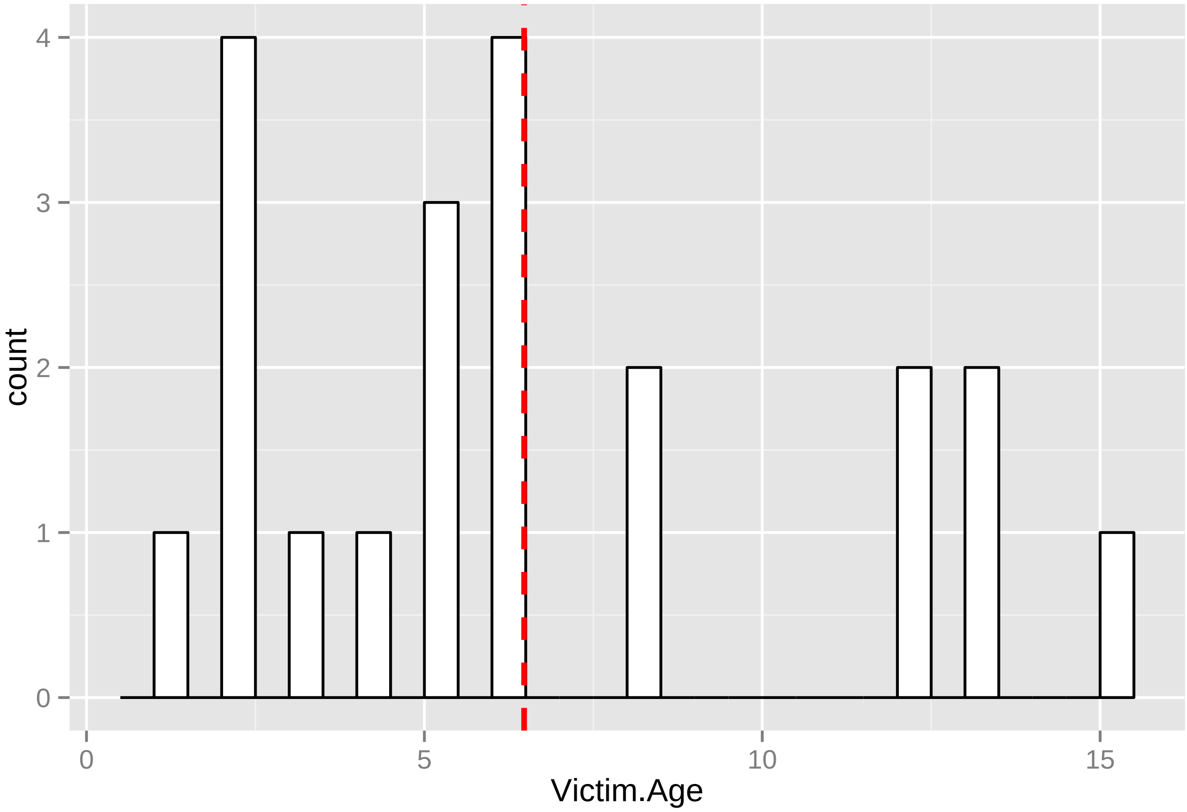
We see higher counts in the Southeast and South Central, as we saw in the geographic plot, and the father appears to be the most common offender across all regions.

```
ggplot(data, aes(x=Relationship, fill=Relate.Group)) +
  geom_histogram(binwidth=.5, colour="black") + facet_grid(Region ~ .) +
    theme(axis.text.x = element_text(angle=90))
```

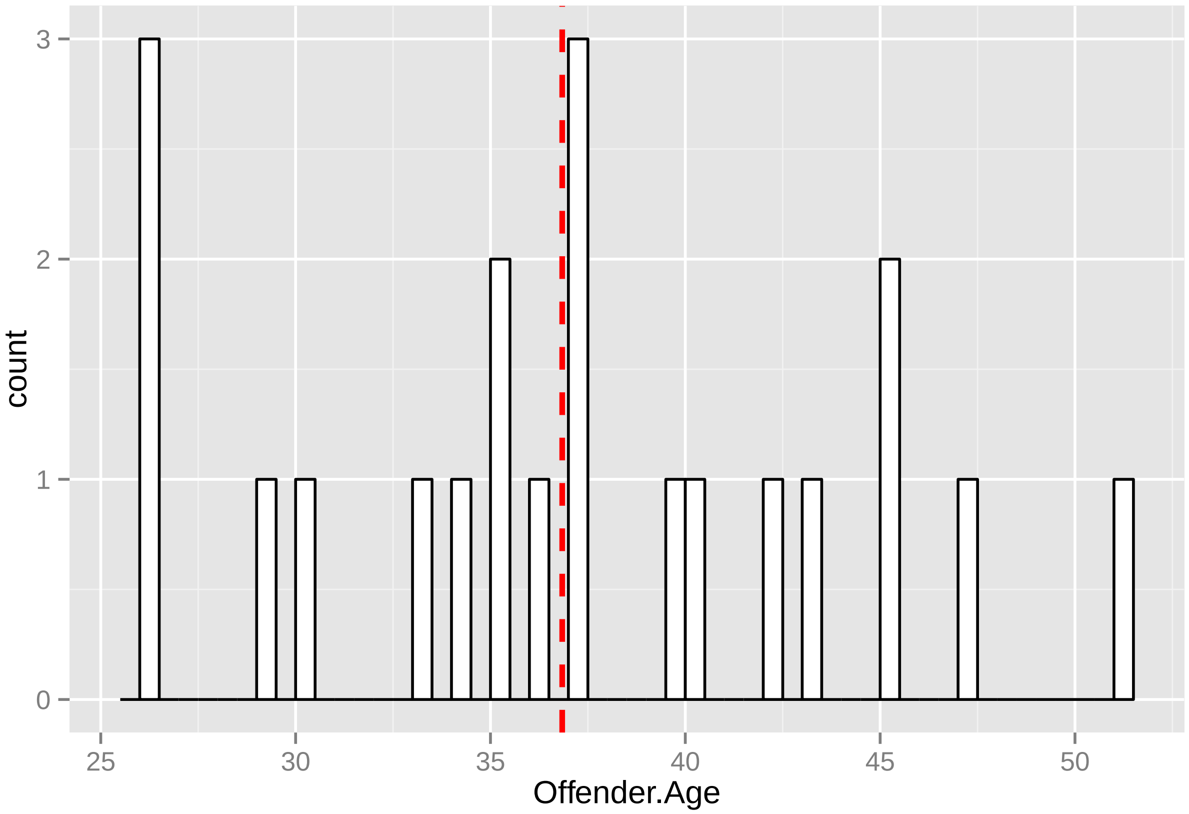


The age of the victims is spread from 1 to 1, with a mean of 6. The age of the offenders range from 26 to 51, with a mean of 37

```
ggplot(data, aes(x=Victim.Age)) + geom_histogram(binwidth=.5, colour="black", fill="white") +
  geom_vline(aes(xintercept=mean(Victim.Age, na.rm=T)),   # Ignore NA values for mean
             color="red", linetype="dashed", size=1)
```

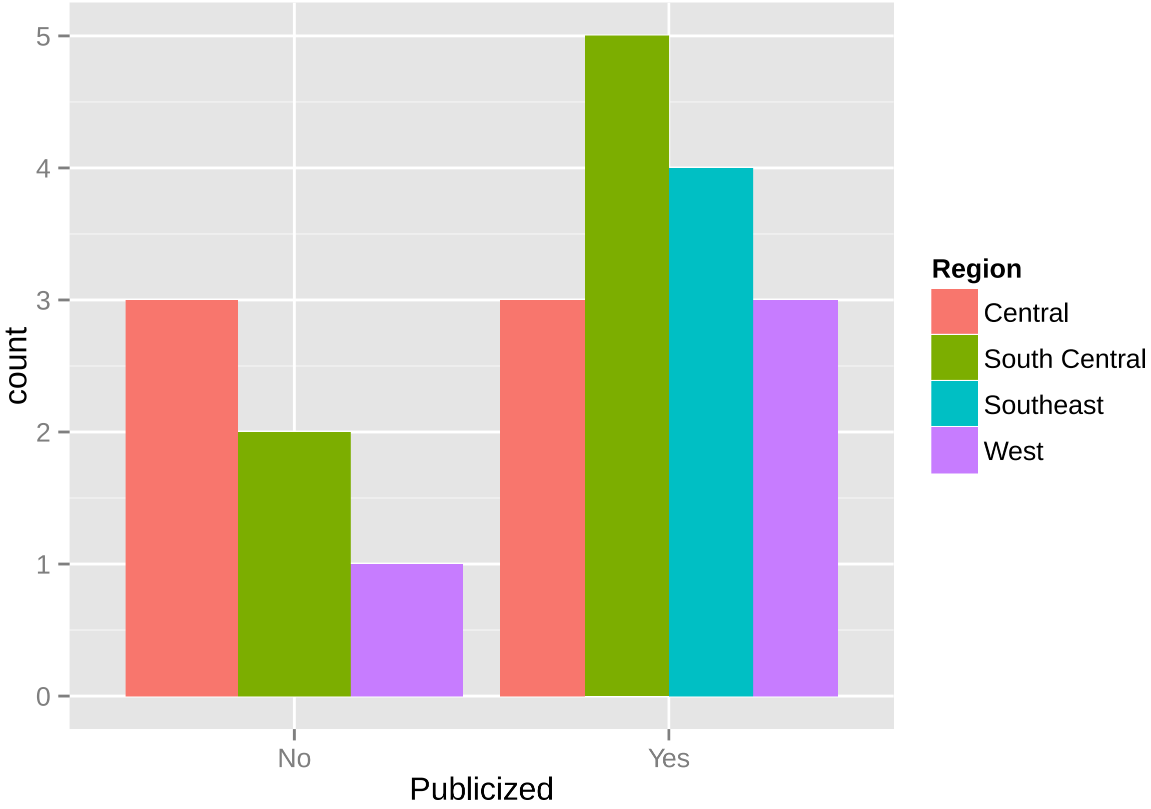


```
ggplot(data, aes(x=Offender.Age)) + geom_histogram(binwidth=.5, colour="black", fill="white") +
  geom_vline(aes(xintercept=mean(Offender.Age, na.rm=T)),   # Ignore NA values for mean
             color="red", linetype="dashed", size=1)
```



The majority of cases are publicized with the highest being in the South Central Region. With the small data set, its difficult to drawn conclusions. If this holds in the final data set, additional sponsor knowledge will be sought to understand why this is the case and the interaction of publicity with the trend toward violence.

```
ggplot(data, aes(x=Publicized, fill=Region)) + geom_histogram(binwidth=.5, position="dodge")
```



Data processing for model building

Several factors were removed from the data set prior to evaluating predictive models. RSO, registered sex offender, only has “No” fields and provides no differentiation. Numerical age groups and detailed city were eliminated in favor of age groups and region. When additional data is available, using the direct age will be considered. Data that is unknown prior to or in the early stages of an abduction (Publicized, Recovery) and the temporary variables (Harm, Relationship) were also removed.

```
# Remove extra attributes and those not known at the time of abduction
data.m <- data[,-which(names(data) %in% c("RSO", "Harm","Date","Victim.Age",
                    "Offender.Age", "Relationship", "Location", "Recovery",
                    "Publicized", "Number.Victims"))]
```

Due to the relatively low rate of occurrence of the target variable, two additional test data sets were created that up-sample the minority target case. Data sets with the target variable as factored and numeric were created to meet the needs of different models.

```
# In SMOTE, target must be factor and last item in dataframe
data.m$target <- as.factor(data.m$target)
data.smote <- SMOTE(target ~ ., data.m, perc.over = 500)
data.smote2 <- SMOTE(target ~ ., data.m, perc.over = 500, perc.under = 125)
smote.output <- as.data.frame(rbind(table(data.m$target),
                table(data.smote$target),table(data.smote2$target)))
rownames(smote.output) <- c("original data","upsampling","upsampling/downsampling")
```

```
# rpartXse in performanceEstimation requires non factored target
data.m.nf <- data.m
data.m.nf$target <- as.numeric(data.m.nf$target)
data.smote.nf <- data.smote
data.smote.nf$target <- as.numeric(data.smote.nf$target)
data.smote2.nf <- data.smote2
data.smote2.nf$target <- as.numeric(data.smote2.nf$target)
```

Summary of target variable in the different data sets:

	0	1
original data	15	6
upsampling	60	36
upsampling/downsampling	37	36

Model building and evaluation

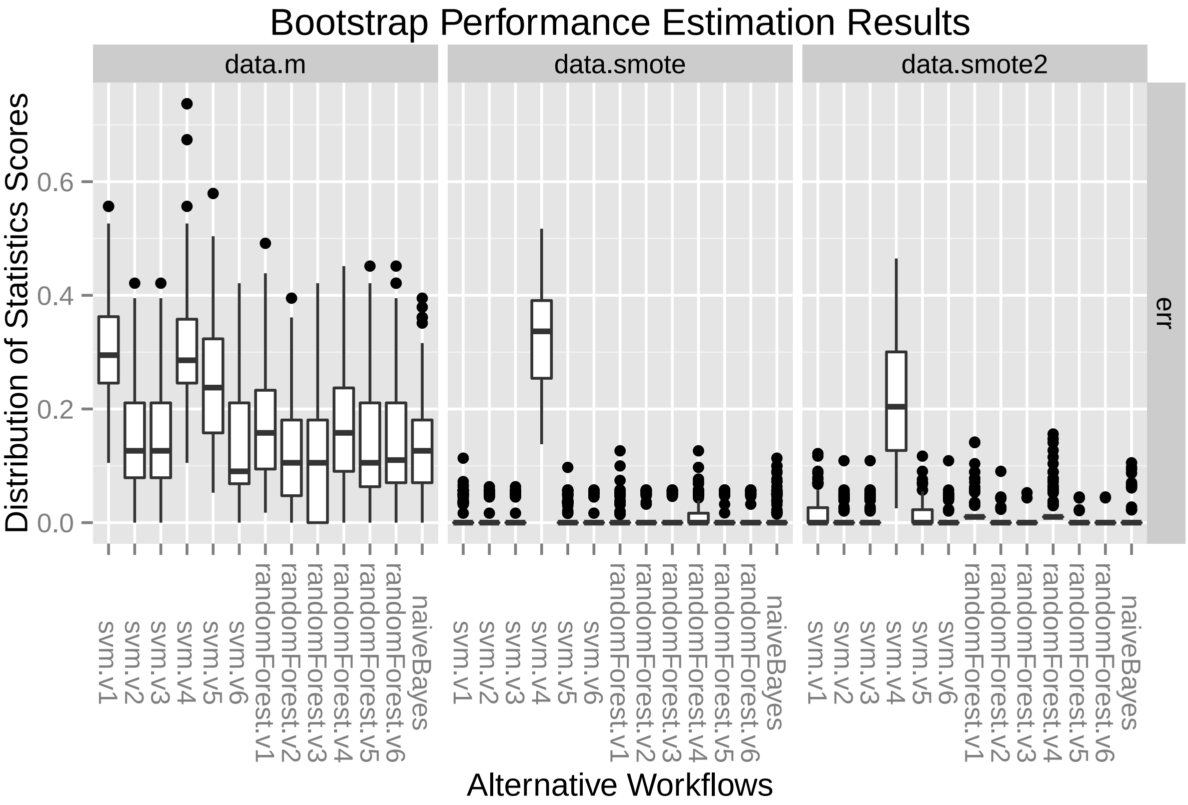
Our initial evaluation used four modeling techniques with assorted options.

- Tree induction (rpart)
- Random Forest
- SVM
- Naive Bayes

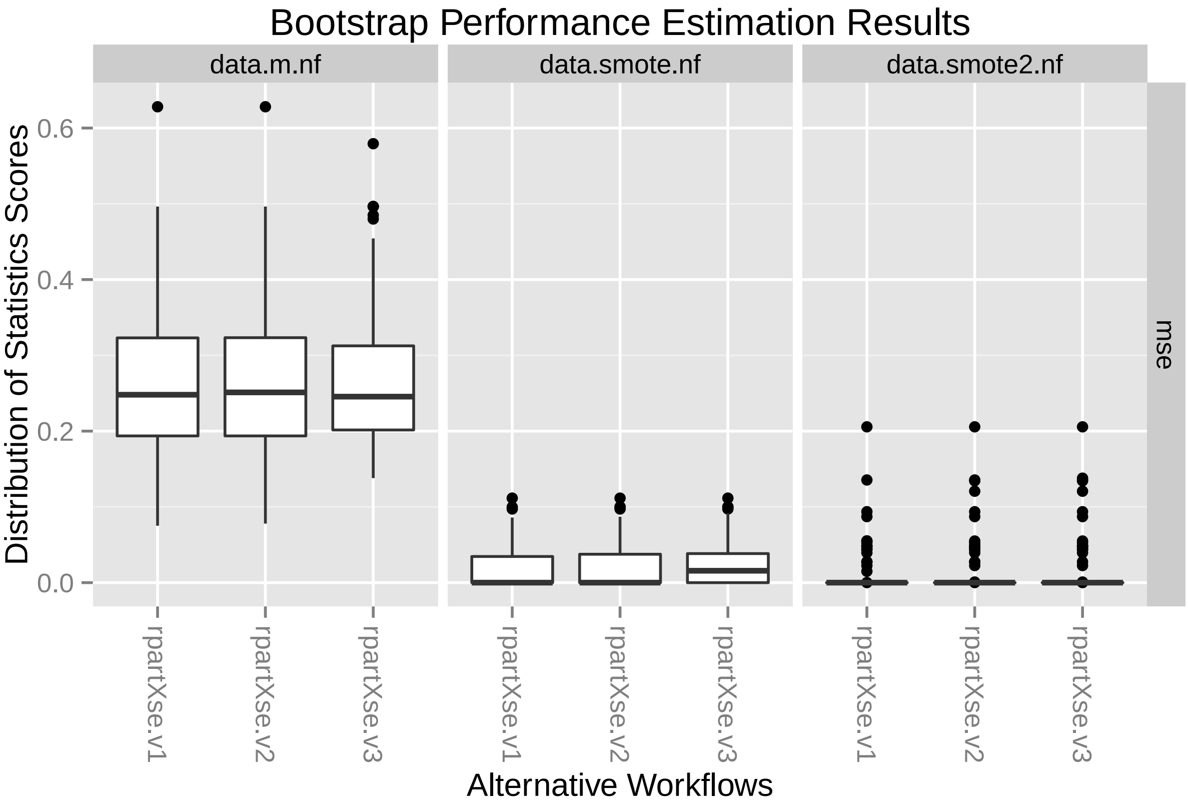
Due to the small sample size, bootstrap sampling was used, which uses random sampling with replacement to create the training data. With the extremely small size of our sample data set, even with bootstrap sampling, the results of different runs tend to be variable. Final interpretations and conclusions on the modeling will be done once we get the full data set next week.

```
# Build and evaluation for rpart (requires non factored target)
res.nf <- performanceEstimation(
  c(PredTask(target ~ ., data.m.nf),PredTask(target ~ ., data.smote.nf),
    PredTask(target ~ ., data.smote2.nf)),
  workflowVariants("standardWF", learner = "rpartXse", learner.pars=list(se=c(0,0.5,1))),
  BootSettings(type=".632", nReps=200))

# Build and evaluation for SVM, NaiveBayes, RandomForest
res <- performanceEstimation(
  c(PredTask(target ~ ., data.m),PredTask(target ~ ., data.smote),
    PredTask(target ~ ., data.smote2)),
  c(workflowVariants("standardWF", learner = "svm",
           learner.pars=list(cost=c(1,10,100), gamma=c(0.1,0.01))),
    workflowVariants("standardWF", learner = "randomForest",
                     learner.pars=list(ntree = c(5,50,200), nodesize = c(2,5))),
    workflowVariants("standardWF", learner = "naiveBayes")),
  BootSettings(type=".632", nReps=200))
```

The models using factored targets vs numerical were run separately. The results of the two sets of runs are shown in the graphs below.



Bootstrap Performance Estimation Results



Bootstrap Performance Estimation Results

The top performers of the two sets are:

```
## $data.m
##             Workflow Estimate
## err randomForest.v3    0.115
##
## $data.smote
##      Workflow Estimate
## err   svm.v6    0.004
##
## $data.smote2
##             Workflow Estimate
## err randomForest.v6        0
```

```
## $data.m.nf
##         Workflow Estimate
## mse rpartXse.v2    0.262
##
## $data.smote.nf
##         Workflow Estimate
## mse rpartXse.v1    0.018
##
## $data.smote2.nf
##         Workflow Estimate
## mse rpartXse.v1    0.005
```

Typically, the top performers are svm.v2, naiveBayes, randomForest.v6. That 200 repetition bootstrap still gives varying results is indicative the the sample size is insufficient. With 20 observations and 10 factors, the models are undoubtedly over-fitting.

We expect to receive additional data next week from our FBI sponsor and will run this through the existing models. Her interest is primarily in identifying risk factors. Models such as svm and randomForests, that may provide better predictions, may not meet her needs.

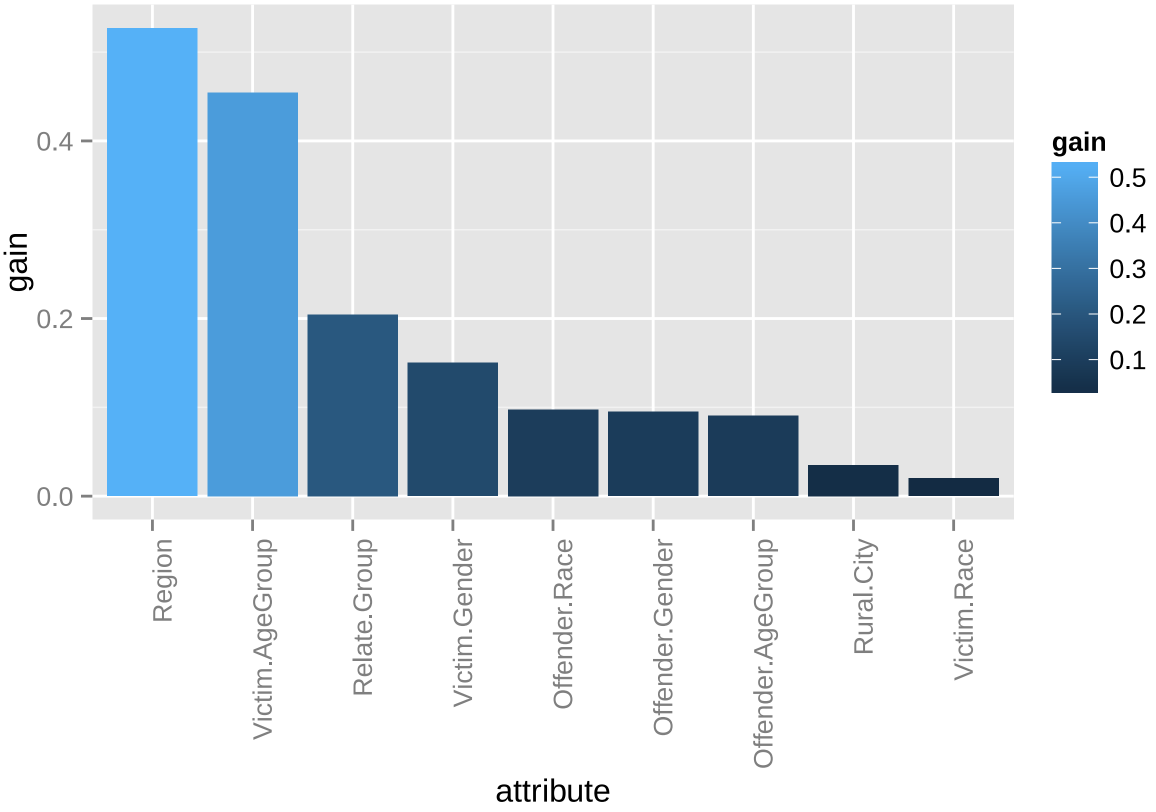
Risk factors using Naive Bayes

In preparation for discussing the data and preliminary results, we ran the Naive Bayes model on the up-sampled data and identified the factors that are providing the most information gain.

```r
library("RWeka")
NB <- make_Weka_classifier("weka/classifiers/bayes/NaiveBayes")
model.weka <- NB(target ~ ., data.smote)
infogain <- as.data.frame(InfoGainAttributeEval(target ~ ., data.smote))
ig <- cbind(rownames(infogain),infogain)
colnames(ig) <- c("attribute","gain")
summary(model.weka)
```

```
##
## === Summary ===
##
## Correctly Classified Instances          96               100      %
## Incorrectly Classified Instances         0                 0      %
## Kappa statistic                          1
## Mean absolute error                      0.0283
## Root mean squared error                  0.0611
## Relative absolute error                  6.0316 %
## Root relative squared error             12.6281 %
## Coverage of cases (0.95 level)         100      %
## Mean rel. region size (0.95 level)      56.25   %
## Total Number of Instances               96
##
## === Confusion Matrix ===
##
##   a  b   <-- classified as
##  60  0 |  a = 0
##   0 36 |  b = 1
```

```
ig$attribute <- factor(ig$attribute, levels = ig[order(-ig$gain),]$attribute)
ggplot(data=ig, aes(x=attribute, y=gain, fill=gain)) + geom_bar(stat="identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



The results are consistent with our sponsor's experience and gut feelings. Younger victims are easier to coerce. Parents that engage in abduction have more tendency to be interested in revenge and thus trend more violently. The regional effect is surprising and bears further investigation. A specific question that she was interested in was whether the location, rural vs city, was important. Our preliminary results indicate that it is not, additional data will confirm/reject this.

Conclusions and next steps

Once we get the additional data, we will quickly run through the analysis and share the results with our sponsor. This preliminary analysis has uncovered a number of interesting questions that can only be answered with the larger data set. We also expect to receive a much larger data sets from National Center for Missing and Exploited Children, NCMEC, which will need to be processed and evaluated to determine if similar modeling is applicable. With all data in hand and in collaboration with our sponsor, we will re-evaluate our choice of target variable and model techniques.