An analysis of Parole Data
Arjun Gupta
Prof. Berk, STAT 474
3/30/15

1.

```
> summary(as.Date(working1$DOB,origin="01-01-1970"))
        Min.      1st Qu.       Median         Mean      3rd Qu.         Max. 
"-048-07-21" "-002-07-30" "0009-02-12" "0006-07-24" "0016-02-22" "0023-12-03" 
> summary(working1)
 Fail          DOB           AdultPriors     JuvenilePriors     iViolCount       Male     
 0:40028   Min.   :-17714   Min.   :  0.00   Min.   : 0.000   Min.   : 0.000   0: 8265  
 1: 7931   1st Qu.:  -904   1st Qu.:  6.00   1st Qu.: 0.000   1st Qu.: 0.000   1:39694  
           Median :  2946   Median : 16.00   Median : 0.000   Median : 0.000            
           Mean   :  2013   Mean   : 27.09   Mean   : 2.833   Mean   : 1.256            
           3rd Qu.:  5512   3rd Qu.: 38.00   3rd Qu.: 1.000   3rd Qu.: 3.000            
           Max.   :  8353   Max.   :150.00   Max.   :92.000   Max.   :16.000            
```

We take a preliminary look at the summary statistics. The first thing to note is that the chance someone fails is relatively small; 84% of those on parole don't have fails. Now to look at predictors. In class, "the mean being less/greater than the median" was sufficient justification for skew. With that definition in mind, we see that the distributions of predictors are skewed—the mean is almost twice the median for Adult Priors, the mean is greater than the 3rd quartile for Juvenile Priors, the mean is greater than the median for iViol count, the ratio of males to females is about 5 to 1, and the while the median age of our sample is 25, the average age is 28. It is important to recognize that skewed distributions do not negatively affect CART results, as CART as is splitting ordinal data to reduce variance—as long as there aren't miscodings, non-normal distributions should not be a problem for CART. Thus, an ordinal variable outlier, like the max of 92 juvenile priors, should not be an issue for CART. Outliers and other data points far to the right of the median should fall to the right of whatever split CART makes for that predictors

Now we consider data quality and potential inconsistencies. We should note that when a minor and an adult each have a "0" for adult priors it could mean different things. For a minor, the "0" might be equivalent to an "NA", (since youth shouldn't be having adult priors) but for an adult it would actually indicate having zero prior charges as an adult. This matters for CART. If there are a large number of observations with zero adult priors that were actually supposed to be NAs, the ordinality of the data will be compromised--CART will be making false splits

When we subset on cases with zero adult priors, only one case out of the nearly 48,000 cases emerges. This alleviates our concern about potential inconsistency between an adult with zero priors and a minor with zero adult priors. But this seems a bit odd, as half our inmates are under 25. When we subset on people born after 1985-01-01 (meaning they would be 18 years old or younger at the beginning of 2003), we have more than 12,000 observations, about a quarter of the data. Except for one observation, every single one of these inmates, less than 18 years old at the beginning of 2003, is listed as having adult priors. In fact, the median adult priors for subjects in our "under 18 "sample is 9. We briefly consider that unbeknownst to us, Philadelphia may be an anomaly, and the age of criminal responsibility may be 17 (which according to the Juvenile Law Center http://www.jlc.org/news-room/media-resources/glossary is possible). Even then, subsetting on being born after 01-01-1986 yields nearly 10,000 observations with adult priors, again with the median of adult priors for subjects *17 or under* being 9. Also, criminals on parole usually have already served some time behind bars, so there are a significant number of minors convicted of adult charges.

One possible explanation is that the crimes these individuals under 18 committed were severe enough to be treated as adult charges. Or there is other definition or law regarding adult priors that makes this phenomenon possible that we are just not aware of.

At any rate, throwing out between a fifth and a quarter of the data is a drastic thing to do without some consultation with the provider of the data first. Since I can't do the proper due diligence within the scope of this paper, and we have not talked about protocols for striking large portions of data, I am assuming there is some other definition of adult charge that makes it possible for a minor to have one. Thus, we will proceed. If this were a real project I would again talk to the provider of the data to understand what's going on and make the decision to strike from there. But for this analysis I will keep them.

Another problem: duplicates in the data. We investigate with the *duplicated* command in R and find that 4150 observations are exact duplicates of other observations. The likelihood that 2 inmates on parole in the Philadelphia parole system have the exact same values for every single predictor is extremely small. This is could be an issue, when 8.6% percent of our data could cause CART to make splits it shouldn't.

On this issue I am inclined to strike these observations. Keeping 91.4% of the data does not seem bad. It does not seem prudent to run CART on a dataset with nearly 9% false signal, and almost 44,000 observations should still be plenty. Again, in real life I would talk with the Parole Office and based on that conversation make my decision. But for this analysis we strike them.

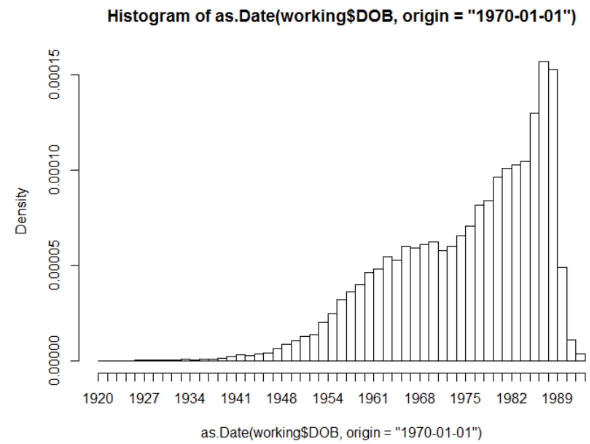
A look at the summary statistics for our final working dataset:

```
> summary(working)
 Fail            DOB           AdultPriors      JuvenilePriors      iViolCount        Male
 0:36445   Min.   :-17714   Min.   :  0.0   Min.   : 0.000   Min.   : 0.000   0: 7665
 1: 7364   1st Qu.: -1032   1st Qu.:  5.0   1st Qu.: 0.000   1st Qu.: 0.000   1:36144
           Median :  2860   Median : 15.0   Median : 0.000   Median : 0.000
           Mean   :  1930   Mean   : 25.4   Mean   : 2.783   Mean   : 1.158
           3rd Qu.:  5453   3rd Qu.: 35.0   3rd Qu.: 1.000   3rd Qu.: 2.000
           Max.   :  8353   Max.   :150.0   Max.   :92.000   Max.   :16.000
```

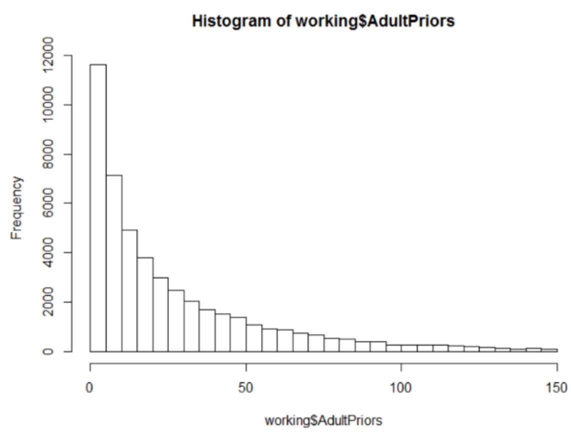
The distributions for the continuous predictors changed little, a good sign. Fails still make up 16% of the sample, and there are still many more males than females. It is also worth noting that now the number of inmates who are 17 and under with adult priors is down to 8745, and the number under 18 is still above 10,000.

Thus we see that this cohort of offenders is relatively young (median age of 25), overwhelmingly male (about 83%), committed most of their crimes when they were considered adults (as juvenile priors is very low), and most are not on parole for violent crimes (as the median iViolCount is 0, and 75% of inmates had 3 or less).

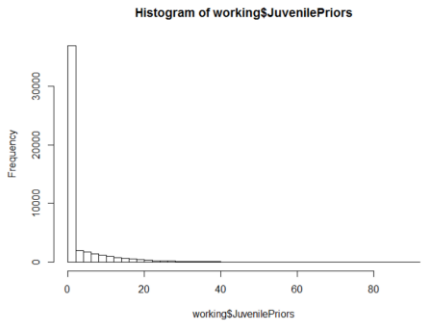
2. Date of Birth

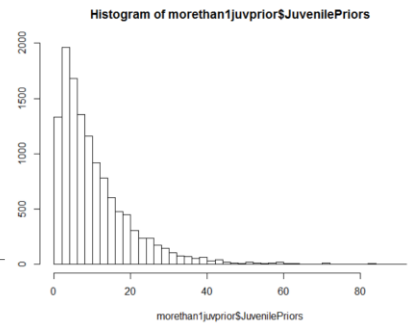


There is a wide range of ages, from 11 years old to a little over 80 years old. But, as one might expect, the distribution of age is left-skewed (the bulk of criminals are younger; the older ones are outliers), with half our inmates are under 25 years of age. This makes sense, as guest lecturer Jordan Hyatt said that individuals with the greatest propensity towards crime are young males. He also said individuals' propensity towards crime drops dramatically after 40, and we see that the majority of the mass in the histogram is to the right of 1963.

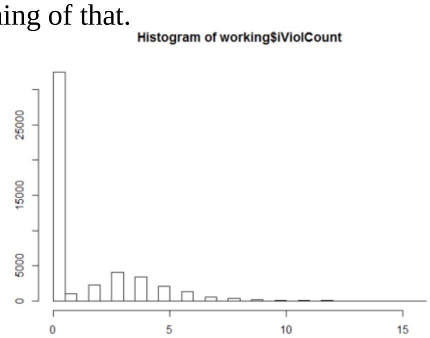


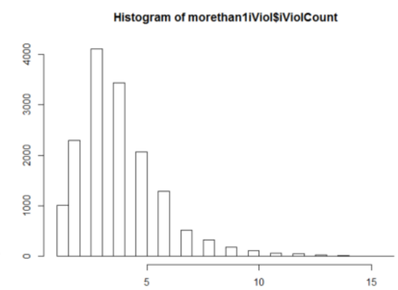
Remembering that one crime usually carries multiple charges makes the distribution of adult priors plausible. We saw that the median number of adult priors is 16 and the mean is 27—quite a right skew The histogram confirms this notion. Very few people have more than 50 prior charges as an adult. What is interesting to see is the cluster of values at 150. When we subset on people with 150 adult priors, we find 25 observations. The fact that there are so many at the max amount makes it seem possible that there might have been an arbitrary cut off at 150 when reporting adult prior data. If this is the case, it would throw CART off, as we have again compromised the ordinality of the data. On the positive side, 25 observations out of 47,959 is small. We will proceed, for it is possible that several people each had 150 priors and not more.





The vast majority of our inmates had a clean record as juveniles with (what appears to be) a small portion of the population having more. It turns out that 25% of inmates have at least one juvenile prior. When we subset on these individuals, we find that the median juvenile priors for them is actually 8. Perhaps CART will make something of that.





With a median of zero and a 3rd quartile of 3, 75% of our inmates have 3 or less charges of violent crime that led to the current conviction. We see that the vast majority of our criminals are generally not violent people.

This examination of the histograms again reinforces that our criminals are mostly young, overwhelmingly male, not very violent, and have many more adult charges than juvenile ones.

3. Here is the table comparing the proportions of fails and males. Gender is on the top row:

	0	1
0	0.15499098	0.67691570
1	0.01997306	0.14812025

The first thing that jumps out is the portion of women that fail--only 2%. As expected, men that fail are 15% of the population, and since our sample is mostly men and we have relatively few fails, 2/3 of our data are men who don't fail. From this we might guess that CART will classify women in the zero category since they don't usually fail.

And below are the summary statistics for each numerical predictor subsetted by fail (1) versus non-fail (0):

```
> tapply(working$DOB,working$Fail,summary)
$`0`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -17710   -1284    2565    1696    5241    8353

$`1`
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-16300.0    799.2   4121.0   3089.0   6102.0   8245.0

> tapply(working$AdultPriors,working$Fail,summary)
$`0`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    5.00   14.00   24.33   34.00  150.00

$`1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    8.00   20.00   30.69   44.00  150.00
```

```
> tapply(working$JuvenilePriors,working$Fail,summary)
$`0`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   2.296   0.000  92.000

$`1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   5.192   8.000  84.000

> tapply(working$iViolCount,working$Fail,summary)
$`0`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   1.116   2.000  16.000

$`1`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   1.365   3.000  14.000
```

We see that across the board those who fail are, on average, younger, have more adult and juvenile priors, and more violent counts of crime. The 50th percentile of the "fail" subset is the same or higher (or younger) across the predictors as well. This yields exactly what we'd expect, that those who are younger and have more of a criminal past are more likely to fail. Will these differences in means be enough for CART to get beyond the fact that fails are still relatively rare? We shall see.

4. The data is divided equally and randomly into thirds. Using the default cost ratio of false negatives and false positives being equal, we run the CART analysis on our training data, and find it does not get beyond the root node. When we used the default costs, since there are relatively few fails, it appears that there are not enough differences between the fails and non-fails for CART to make splits that do "better". That is, CART can't find anything to split on that would group the observations into nodes with less heterogeneity than the one already present.

The confusion table for how the resulting "decision boundary" performs on the evaluation data fitted values are on top:

```
   fitted1
        0     1
  0 12170     0
  1  2433     0
```

Not getting beyond the root node means CART classified all observations by, as phrased in class, "majority vote." Since most criminals don't fail on parole, it classified everyone as a 0, yielding 2433 false negatives. We examine the use errors and model errors:
Model error followed by use error:

```
  fitted1
    0 1
  0 1 0
  1 1 0
```

```
  fitted1
            0 1
  0 0.8358372
  1 0.1641628
```

The use error reflects the marginal distribution of fails in the population. This makes sense, because when it found nothing better to split on, CART classified all observations as zeros. Even when we change the minimum bucket size to 10, it is not enough for CART to get beyond the root node. With a 1:1 cost ratio there is not enough heterogeneity in the sample to create subnodes that are more homogenous.

5. Using the given "New Prior" function, we tune the cost of misclassifying a failure (a false negative) to multiple times that of the cost of misclassifying a non-failure (a false positive) by increasing the prior for failure. Since, for our purposes, the prior is the marginal distribution of fails in the population, this means we are artificially "increasing" the number of fails.

CART uses the Gini index as its default loss function, p*(1-p), where p is the probability of being in the positive class **(**Berk, *Statistical Learning from a Regression perspective*, pg. 114**)**. This function is maximized at p=.5. When we tune the cost ratio to 4:1, the prior becomes .44. This is much closer to the maximal value—there is much more “impurity” of 1s and 0s in the population, so now CART has more to chances to find splits on different variables that will increase homogeneity in the subsequent nodes.

It is makes sense that the damage to property, lives, economic activity, and social cohesion of a parolee committing a crime is a good bit more costly than spending extra resources on someone who doesn’t need them. Both Jordan Hyatt and Professor Berk have offhandedly mentioned about a 5:1 cost ratio, so we’ll shoot for between a 4:1 and 7:1 empirical cost ratio in the evaluation and test data

Using a prior of .44 yields the following confusion table for the evaluation data:

```
  fitted2
       0    1
0 9635 2535
1 1452  981
```

We have made false negatives more costly than false positives, but by less than 2:1. We try the using a tuning ratio of 8:1. This yields a prior for the positive class of .61, and the following confusion table for evaluation data:

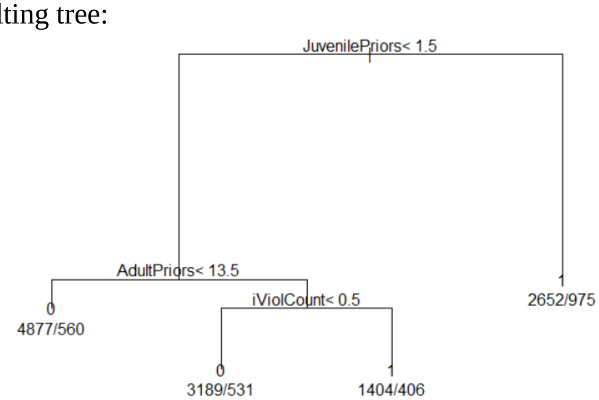
```
  fitted3
       0    1
0 3702 8468
1  318 2115
```

The empirical cost ratio of false negatives to false positives in the evaluation data is now over 26:1. Too much!

The tuning cost ratio of 5:1 yields a prior for the positive class of .497. and the following confusion table:

```
  fitted4
       0    1
0 8186 3984
1 1069 1364
```

The empirical cost ratio in the evaluation data is 3.7:1—close to what we wanted originally. However, when we examine the resulting tree:



There is no mention of Date of Birth, something most policymakers and criminologists will likely say matters. When we remember that these people are the end users of the forecasts, it is difficult to imagine them accepting decision boundaries that don’t take age into account.

So we try the tuning parameter cost ratio of 5.2:1 and a prior of .512. This yields the following confusion table:

```
  fitted5
       0    1
0 6771 5399
1  807 1626
```

This is an empirical cost ratio of a little under 7:1, again within our bounds. We examine the model (left) and use (right) errors:

fitted5

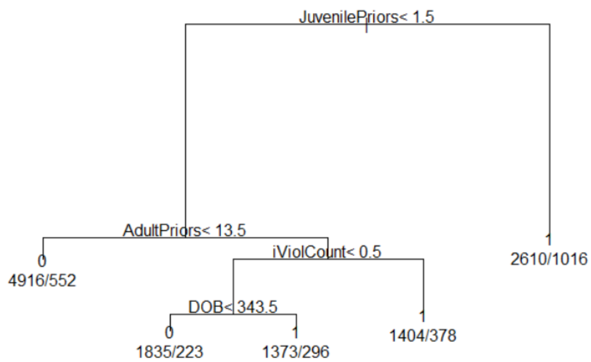
	0	1
0	0.5563681	0.4436319
1	0.3316893	0.6683107

fitted5

	0	1
0	0.8935075	0.7685409
1	0.1064925	0.2314591

The additional weighting of false negatives comes at a cost. From the table on the right, nearly 80% of the people we say will commit crimes actually don't, and given the truth in any category our model never finds it more than 2/3 of the time (from the table on the left). Our overall model error is a shoddy (807+5399)/14603=.43. But, the policymakers will be happy to know that we only have a 10% error rate when we predict someone will not commit a crime. While the costs from the weighting in terms of both model metrics and additional resources devoted to those who don't need them are high, the streets will be safe when using these decision boundaries.

Examining the tree:



It is important to note that regression trees are not to normally used for inference for reasons we will discuss. However this question calls for interpretation, and such interpretation can be instructive.

According to this tree, those with multiple juvenile priors are bad apples with regards to parole, as anyone with greater than 1.5 juvenile priors is immediately classified as a positive. It makes sense that adult priors is next, for one would expect that past criminal tendencies tend to predict future ones. It also makes sense that lots of adult priors and violence also put someone in the positive class—it costs a lot to have violent criminals fail on parole. Finally, it's interesting to note age is deepest in the tree, and does not have the predictive power we expected when we were looking at the initial histograms of the data.

6. We use these decision boundaries to construct a confusion table on the test data:

fittedtest

	0	1
0	6802	5335
1	764	1702

Followed by the model error (right) and use error (left)

fittedtest

	0	1
0	0.5604350	0.4395650
1	0.3098135	0.6901865

fittedtest

	0	1
0	0.8990219	0.7581356
1	0.1009781	0.2418644

The empirical cost ratio on test data is still close to 7:1. Our false negative rate is still 10%, and the false positive rate fluctuates to 75.8%. The similarity in the cost ratios and use error between our confusion tables from evaluation and test data is good. It suggests that we have decision boundaries that will have low variance when forecasting sample to sample. We can expect around 10% error when forecasting fails, exactly what we assume policymakers likely care about most—public safety.

7. We have discussed how well the tree performed in the test data set above, but how well will it perform in the future? We can assume that the forecasting skill demonstrated in the test data will apply to future data sets *if we assume that future data will also be random samples from the same population from which this data came* (Berk 139). This seems a reasonable assumption. Professor Berk himself remarks "in the world of parole, things have not changed much" (Berk Parole Assignment). Furthermore, it seems safe to say that criminals don't change much from year to year. But, for example, the parole office might hesitate to use this model to forecast parole violations during a recession, when perhaps the economic trend and general panic national economy have the potential to influence the nature of the criminal population on which they are forecasting.

What sort of inference can we do from these trees? It is prudent to first recognize the mindset with which we are doing this inference. We are not trying to simply describe the data at hand but forecast on future samples. If there is even is a true response model f(X), it would be difficult to claim we know it and are simply trying to estimate it—for one, there is no way we have included all the possible predictors associated with failing on parole or not. Socioeconomic status, education level, drug use, and many other things we haven't accounted for likely influence parole outcomes. We are also assuming there is no measurement error, and from the inconsistencies we found when initially examining the data it is hard to make that claim. Thus, we are necessarily in the wrong model approach, and our intent is "to construct a best guess of the values of a set of conditional proportions in a population" (Berk 135). We are necessarily also constrained to making inferences specifically when we consider false negatives more costly than false negatives. If these weights were reversed, our trees could look very different, and of course these inferences would not generalize.

With that in mind, what we learn from this analysis is that if we don't weight false negatives more than 4 times as costly as false positives, age *may* not matter as much as we might expect. It only appears in the tree when we increase the prior of fails to >.51, and even then it is the last subnode. Juvenile priors and adult priors seem to matter the most and 2nd most, respectively. This holds true when our empirical cost ratio in the test data is 3:1 all to the way to when our cost ratio is 26:1.

It is important to recognize these inferences are limited--trees are normally not used for inference. There is no formal mathematical reasoning that states CART will find a correct function even if it were given a random sample with no measurement error and all the necessary predictors (Berk 149-150). If the functional form of tree is wrong, it follows that there is bias which throws off our inference.

CART trees are also very unstable. Tweaking minimum bucket size and tuning parameters can drastically change the structure. For example, when we changed the prior of Fail to .497, it yielded a tree that didn't include Date of Birth, yet a tree using the prior of .512 did. Two of our terminal nodes are still fairly heterogenous:

```
    21) DOB>=343.5 1669  806.0933 1 (0.4730849 0.5269151) *
  11) iViolCount>=0.5 1782  824.2935 1 (0.4182474 0.5817526) *
```

Had a few observations in each of these nodes had a different outcome (and in a future draw of the data, there is a fair chance they might), the node classifications could have been different.

All in all, we have many reasons that we don't do inference directly on trees produced by CART. We might very tentatively say that if we take misclassifying a failure on parole to be more costly than misclassifying a non-failure, juvenile priors and adult priors are the predictors (in that order) explain the much, and age only explains very little variation in the response. However, given all the issues detailed in the last few paragraphs, we would be hesitant to make that statement.

On the predictive side, when faced with a relatively homogenous distribution, if we (or policymakers) decide that one type of error is significantly more costly than another, our CART tree forecasts that one type of class with about 10% error (while doing very poorly on all other use, model, and overall error metrics) and with little variance from random sample to random sample. How limiting this is in the future depends on if criminologists have reason to believe the population of criminals is changing over time and if the parole office for some reason sends cases systematically rather than randomly. Thus, the CART model should be implemented with a careful eye and the consultation of subject-matter experts.