Chicago Crime Predictive Analysis

Jen-Li Chen

June 2019

1 Introduction

For the Chicago Crime data set since 2001 until now, I did the variable selection based on the following points which I am most interested in:

- ***Is it local?*** This can be analyzed through the variables: Zip code, Blocks, Police.Districts, Police.Beats,Community.Areas. However in this case, I found before making sense of Zip code, Blocks, Police.District and Police.Beats, it may require more time to do feature engineering and data cleaning. I excluded these variables at this moment due to time restriction.
- ***What locations are more prone for a crime to happen?*** For this perspective, the variables: Location and Location.Description are included in the model. I am interested in usually what kind of location may be easier to be the target places of what sort of crimes.
- ***Is it domestic?*** The variable, Domestic, can help us have an initial understanding of the nature of the crime. Was the crime happened indoor or outdoor? This may reveal whether the criminals and victims knew one another or not.
- ***Arrest or not?*** The variable, Date, lets us understand when the crime had happened? Was it near from now or long ago? The variable, Arrest, gave us the picture that have the crime offenders been arrested or not? This can lead to the analysis such as what scale of the crime would usually more easily to be cracked? How long would it become harder to arrest the criminal? We usually have the sense that the longer the crime happened before now, the harder for the case to be cracked.

2 Basic Concept

2.1 Data Cleaning

My major data cleaning and data analysis procedures are introduced as below:

Firstly, our interested target is the Primary.Type, which during the analysis I further converted it to 6 levels, better for building models: ”Theft”, ”Battery”, ”Criminal Damage”, ”Narcotics”, ”Assault” and ”Others”.

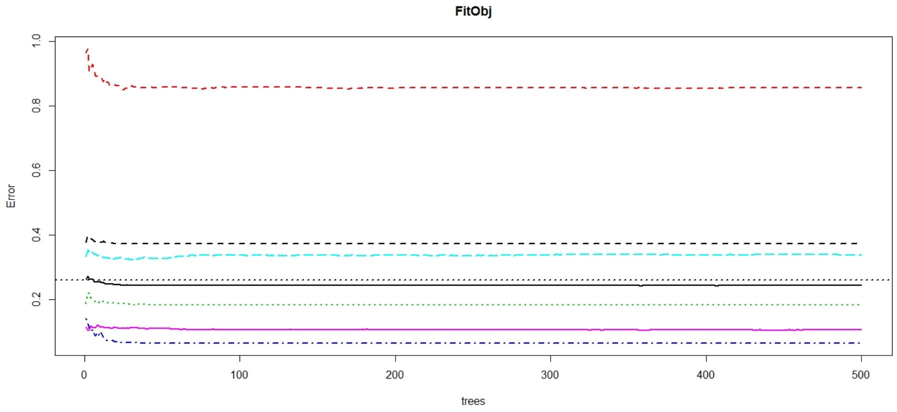
Secondly, I categorized the variable Description to 7 levels: ”Simple”, ”$500 and under”, ”Domestic Battery”, ”Vehicle”, ”Property”, ”Over $500”, and ”Others”.

For the variable Location.Description, I reduced its levels to 6: ”Street”, ”Residence”, ”Apartment”, ”Sidewalk”, ”Parking Lot”, and ”Others”.

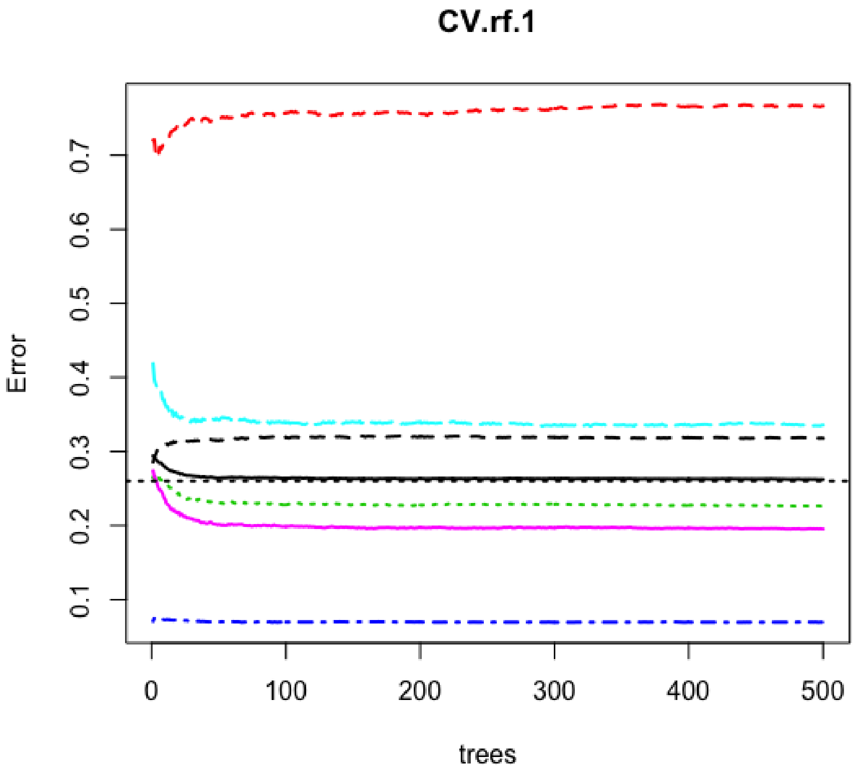
Other variables I chose are Date, Arrest, and Domestic. As for the variable, Date, I converted its data type to numeric.

2.2 Methodology

I used Random Forest with 3-fold cross validation 500 trees to train the model. The error rate is about 24.3% without cross validation and 26.22% with 3-fold cross validation. Each split used 5 variables. The plot of the error rate with different number of trees is as the plot below, which also presented that 500 trees gave us the lowest error rate. This is a decent model to adopt for this case.

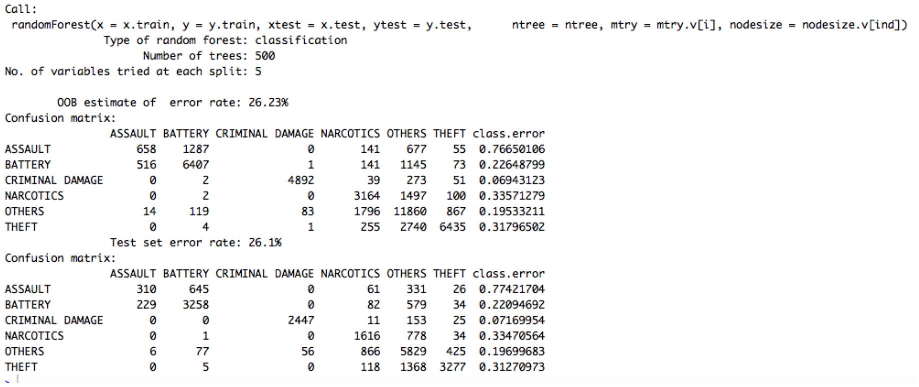


The random forest with 3-fold cross validation presented the error versus number of trees as below. Although random forest with cross validation usually tend to give us higher error rate, in order to avoid over-fitting, cross validation is highly recommended.



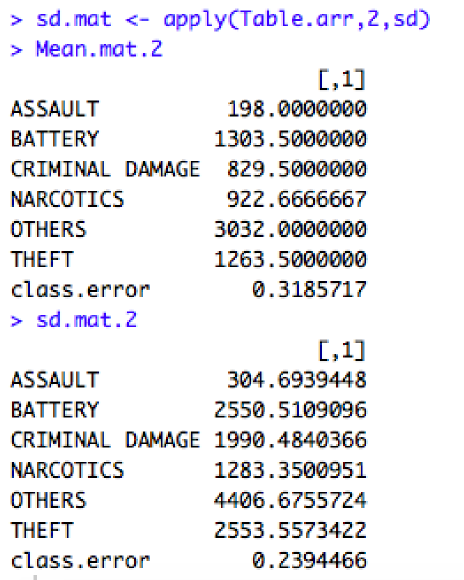
The input variables I chose for random forest model are: Date, Arrest, Domestic, Description of the crime, Local Description. The confusion matrix for random forest with 3-fold cross validation and its mean and standard deviation matrices are as below, showing how much the classification error is for each input variable. We can see "Criminal Damage" has the lowest error rate while "Assault" has the highest error rate.

The model of random forest with 3-fold cross validation is as below:



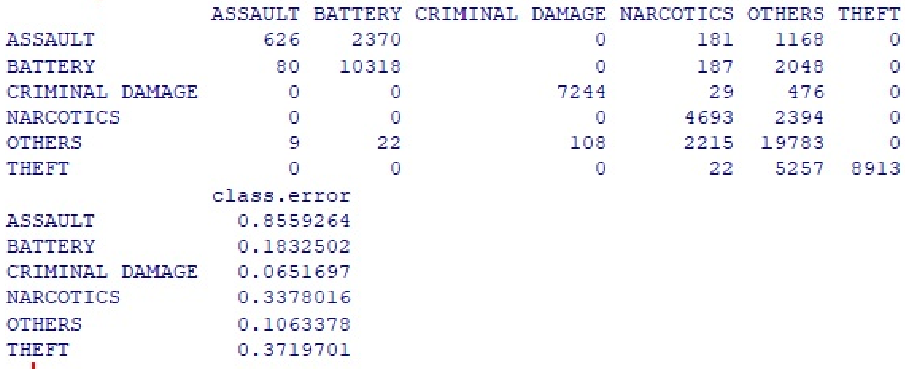
```
Call:
 randomForest(x = x.train, y = y.train, xtest = x.test, ytest = y.test,      ntree = ntree, mtry = mtry.v[i], nodesize = nodesize.v[ind])
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 5

        OOB estimate of  error rate: 26.23%
Confusion matrix:
                ASSAULT BATTERY CRIMINAL DAMAGE NARCOTICS OTHERS THEFT class.error
ASSAULT             658    1287               0       141    677    55  0.76650106
BATTERY             516    6407               1       141   1145    73  0.22648799
CRIMINAL DAMAGE       0       2            4892        39    273    51  0.06943123
NARCOTICS             0       2               0      3164   1497   100  0.33571279
OTHERS               14     119              83      1796  11860   867  0.19533211
THEFT                 0       4               1       255   2740  6435  0.31796502
                Test set error rate: 26.1%
Confusion matrix:
                ASSAULT BATTERY CRIMINAL DAMAGE NARCOTICS OTHERS THEFT class.error
ASSAULT             310     645               0        61    331    26  0.77421704
BATTERY             229    3258               0        82    579    34  0.22094692
CRIMINAL DAMAGE       0       0            2447        11    153    25  0.07169954
NARCOTICS             0       1               0      1616    778    34  0.33470564
OTHERS                6      77              56       866   5829   425  0.19699683
THEFT                 0       5               0       118   1368  3277  0.31270973
>
```



```
> sd.mat <- apply(Table.arr,2,sd)
> Mean.mat.2
                        [,1]
ASSAULT          198.0000000
BATTERY         1303.5000000
CRIMINAL DAMAGE  829.5000000
NARCOTICS        922.6666667
OTHERS          3032.0000000
THEFT           1263.5000000
class.error        0.3185717
> sd.mat.2
                        [,1]
ASSAULT          304.6939448
BATTERY         2550.5109096
CRIMINAL DAMAGE 1990.4840366
NARCOTICS       1283.3500951
OTHERS          4406.6755724
THEFT           2553.5573422
class.error        0.2394466
```

The confusion matrix for random forest without cross validation is as below.



```
                ASSAULT BATTERY CRIMINAL DAMAGE NARCOTICS OTHERS THEFT
ASSAULT             626    2370               0       181   1168     0
BATTERY              80   10318               0       187   2048     0
CRIMINAL DAMAGE       0       0            7244        29    476     0
NARCOTICS             0       0               0      4693   2394     0
OTHERS                9      22             108      2215  19783     0
THEFT                 0       0               0        22   5257  8913
                class.error
ASSAULT           0.8559264
BATTERY           0.1832502
CRIMINAL DAMAGE   0.0651697
NARCOTICS         0.3378016
OTHERS            0.1063378
THEFT             0.3719701
```

In addition, I tried K-nearest neighbors model (KNN) with cross validation. However, in this case, I encountered ”there are too many ties” error. The potential reason lies on that I converted categorical variables into numbers and the classification are all discrete classes. This algorithm cannot handle ties properly.

Ultimately, random forest was proven to be more appropriate for this case.

3 Advantage

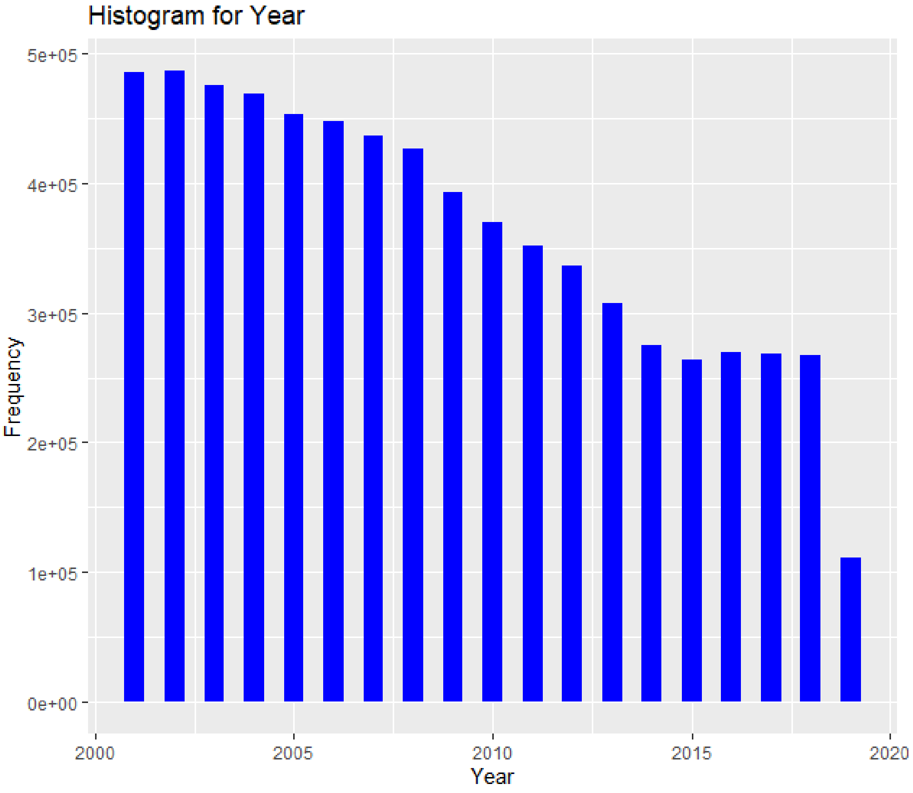
Random Forest model and KNN are both unsupervised and non-parametric models. The advantages of unsupervised models are their flexibility, and they work pretty well without the presumption of the distribution of the data.

4 Drawback

The KNN model is based on the distance between variables, Thus, all the variables in the model have to be numeric. It would not make sense to some data sets or some variables, such as the Zip Code, Police.District, Police.Beat and Community.Areas here. We need to do more sophisticating feature engineering for such variables.

5 Key Finding

- The number of crime is drastically decreasing over time, but in 2015 the progress was obviously stagnant. From 2018 to 2019, the count of cases dropped suddenly to the historical low point since 2000. The plot is as FIGURE below. This finding does not necessarily mean that after 2015 there are fewer crimes than any time before or after, it may also due to recording mistakes, new system or new staff conducting the research, or any other more complicated reasons.



- From the random forest model, the "Criminal Damage" was presented as the most influential classification. The random forest model with 500 trees gave us about 24.3% error rate without cross validation and about 26.22% error rate with 3-fold cross validation. I also would consider checking with VIF and Cook's distance to better reveal the relationship between independent variables and response variable and the correlation among the independent variables.

6 Reflection

If I were given more time, I would like to try the following:

- Perform more feature engineering to reveal in-depth information behind each variable, and how they interact with the target variable. For instance, Police.District, Police.Beat, and Community.Area. I would like to understand more about their differences with one another and their effects on preventing crimes or cracking crimes.
- Decision Tree: I would like to see the difference between random forest and traditional decision tree.
- Gradient Boosting Machine: It uses similar idea as random forest, and I would like to see whether this model can perform better.
- Other methods such as conditional inference trees and SVM with linear kernel methods. I did try these two for this case but my equipment could not afford enough memory to finish these two by the due time.
- Include higher percentage of data in my training data set. Due to R memory limit and time restriction, I used 1% of the whole data set as my training set.
- Lastly, I wish I would have been able to have better equipment to run the models more efficiently and quickly. More professional equipment or utilizing cloud-computing would effectively reduce the waiting time and help debugging quicker and easier.