ECON900 Final Report

Guo Li

In this project, I will try to make predictions on crime types in Chicago based on historical data obtained from Chicago Police database.

1. Data Description and Selection

The whole dataset downloaded from Chicago Police database has 30 columns and more than 6 million rows that include every reported crime incident in Chicago from 2001 to 2019.
My first step is to find out which columns could be used for my prediction model. In the 30 columns of subjects, many of them have duplicate meaning. For example, the column "Location" is just a combination for columns "Latitude" and "Longitude". Many columns would not be useful for this project such as "ID" and "Case Number". Moreover, since this project is to predict the primary type of crimes, I want to less independent variables and make the model simpler and more useful.

After some reading on crimes prediction literature, I decided to focus on location related data to make the predictions.
In this dataset, locations are presented in many subjects including "Beat", "District", "Ward", "Police Beats" etc. While they all give some information on location, many are duplicated in terms of describing a wide crime location. I also found a lot of rows that have missing values in columns like "Wards", "Boundaries - ZIP Codes" etc. If I drop all of them, the dataset size shrink significantly. After some careful consideration, I decided to include "Location Description", "District" and "Community Area" as variables for location. Note that "District" and "Community Area" are represented in numbers and each number represents a uniquely coded district or community. The location information can be found on City of Chicago data website.

In addition to the location subjects, I also included “Arrest” and “Domestic” columns in my analysis, because they are related to both location and crime types.

2. Models and Results

Three models are used for to compare prediction results.

a. Random Forest Classifier

The first model is a normal random forest classifier model just like the ones we used in class. One significant characteristic for this dataset is that all variables are “classes” rather than just “numbers”. So an important step I ran before applying the prediction models is to factorize every variable.

I also tried different “n_estimators” numbers for the model. At first, I tried 100 and my computer eventually crushed after a long wait. But I still want the model to be more precise and I read 64 can be a good spot to run. So in the end, the model has n_estimators = 64 and criterion = 'entropy'.

The following picture shows results for accuracy_score, confusion_matrix and classification_report:

```
0.43479066701041374
[[17983  1346    11 ...     0     0     0]
 [ 1148 14371    23 ...     0     0     0]
 [  287   610    32 ...     0     0     0]
 ...
 [    0     0     0 ...     0     0     0]
 [    1     0     0 ...     0     0     0]
 [    1     0     0 ...     0     0     0]]
C:\Users\Guo\AppData\Local\Programs\Python\Python37\lib\site-packages\sklearn\metrics\classi
1143: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in
no predicted samples.
  'precision', 'predicted', average, warn_for)
                                   precision    recall  f1-score   support

                  CRIMINAL DAMAGE       0.24      0.10      0.14    177212
                CRIMINAL TRESPASS       0.41      0.32      0.36     44774
                WEAPONS VIOLATION       0.18      0.00      0.00     16950
                          BATTERY       0.51      0.61      0.55    283261
                            THEFT       0.41      0.73      0.53    329930
                          ROBBERY       0.29      0.07      0.11     58770
               DECEPTIVE PRACTICE       0.59      0.18      0.28     62564
                          ASSAULT       0.20      0.00      0.01     96943
                        NARCOTICS       0.56      0.90      0.69    162251
              MOTOR VEHICLE THEFT       0.26      0.13      0.17     71204
                    OTHER OFFENSE       0.26      0.09      0.14     97006
                         BURGLARY       0.33      0.55      0.41     90178
 INTERFERENCE WITH PUBLIC OFFICER       0.00      0.00      0.00      3752
           PUBLIC PEACE VIOLATION       0.15      0.00      0.00     11223
       OFFENSE INVOLVING CHILDREN       0.38      0.02      0.03     11064
             LIQUOR LAW VIOLATION       0.33      0.10      0.16      3086
                        OBSCENITY       0.00      0.00      0.00       153
CONCEALED CARRY LICENSE VIOLATION       0.33      0.02      0.04        96
                         STALKING       0.00      0.00      0.00       809
                       KIDNAPPING       0.00      0.00      0.00      1375
                         HOMICIDE       0.89      0.29      0.44      2294
              CRIM SEXUAL ASSAULT       0.06      0.00      0.00      6526
                     PROSTITUTION       0.41      0.19      0.26     15142
                      SEX OFFENSE       0.27      0.00      0.00      5842
                            ARSON       0.12      0.00      0.00      2553
                     INTIMIDATION       0.00      0.00      0.00       877
                HUMAN TRAFFICKING       0.00      0.00      0.00        14
                         GAMBLING       0.19      0.01      0.01      3304
                     NON-CRIMINAL       0.00      0.00      0.00        42
           OTHER NARCOTIC VIOLATION       0.00      0.00      0.00        32
                 PUBLIC INDECENCY       0.00      0.00      0.00        41
 NON-CRIMINAL (SUBJECT SPECIFIED)       0.00      0.00      0.00         2
                   NON - CRIMINAL       0.00      0.00      0.00        10
                        RITUALISM       0.00      0.00      0.00         6
```

The overall accuracy score is around 43.47%. I know this score is not particularly high, but compared to an initial model I used to test run different parameters, the increase is quite significant.

In the test run models, I only used a small fraction of the data (about 10,000 rows) and only included location related data. The best accuracy score I got is around 30%. With more entries in the full dataset and the inclusion of "Arrest" and "Domestic", the improvement is significant.

In the classification report, I tend to trust more on the f1-score as this score takes both false positives and false negatives into account. Three crime types stand out in terms of f1-scores: "Battery", "Theft" and "Narcotics". They are also the ones with most support counts (the exception is "Criminal Damage"). Interestingly, "Homicide" stands out with 0.89 precision score. So we have a lot correctly predicted positive observations to the total predicted positive observations in this category.

b. Random Forest Classifier with One Hot Encoding

As stated previously, all of the variables are factorized and eventually we are looking at different categories rather than plain numbers. So I think the popular One Hot Encoding method could help in this situation.

Compared to the first model, the second model converted all variables with One Hot Encoding before being used in the random forest classifier.

Again, the following picture shows results for accuracy_score, confusion_matrix and classification_report:

```
0.4349125176523101
[[17985  1351    11 ...     0     0     0]
 [ 1145 14379    23 ...     0     0     0]
 [  289   611    32 ...     0     0     0]
 ...
 [    0     0     0 ...     0     0     0]
 [    1     0     0 ...     0     0     0]
 [    1     0     0 ...     0     0     0]]
C:\Users\Guo\AppData\Local\Programs\Python\Python37\lib\site-packages\sklearn\metrics\classificatio
1143: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels
no predicted samples.
  'precision', 'predicted', average, warn_for)
                                   precision    recall  f1-score   support

                  CRIMINAL DAMAGE       0.24      0.10      0.14    177212
                CRIMINAL TRESPASS       0.41      0.32      0.36     44774
                WEAPONS VIOLATION       0.19      0.00      0.00     16950
                          BATTERY       0.51      0.61      0.55    283261
                            THEFT       0.41      0.73      0.53    329930
                          ROBBERY       0.29      0.07      0.11     58770
               DECEPTIVE PRACTICE       0.59      0.18      0.28     62564
                          ASSAULT       0.20      0.00      0.01     96943
                        NARCOTICS       0.56      0.90      0.69    162251
              MOTOR VEHICLE THEFT       0.26      0.13      0.17     71204
                    OTHER OFFENSE       0.26      0.09      0.14     97006
                         BURGLARY       0.33      0.55      0.41     90178
 INTERFERENCE WITH PUBLIC OFFICER       0.00      0.00      0.00      3752
            PUBLIC PEACE VIOLATION       0.16      0.00      0.00     11223
        OFFENSE INVOLVING CHILDREN       0.40      0.02      0.03     11064
             LIQUOR LAW VIOLATION       0.34      0.10      0.16      3086
                        OBSCENITY       0.00      0.00      0.00       153
CONCEALED CARRY LICENSE VIOLATION       0.33      0.02      0.04        96
                         STALKING       0.00      0.00      0.00       809
                       KIDNAPPING       0.00      0.00      0.00      1375
                         HOMICIDE       0.94      0.29      0.44      2294
              CRIM SEXUAL ASSAULT       0.07      0.00      0.00      6526
                     PROSTITUTION       0.41      0.19      0.26     15142
                      SEX OFFENSE       0.30      0.00      0.00      5842
                            ARSON       0.14      0.00      0.00      2553
                     INTIMIDATION       0.00      0.00      0.00       877
                HUMAN TRAFFICKING       0.00      0.00      0.00        14
                         GAMBLING       0.20      0.01      0.01      3304
                     NON-CRIMINAL       0.00      0.00      0.00        42
          OTHER NARCOTIC VIOLATION       0.00      0.00      0.00        32
                  PUBLIC INDECENCY       0.00      0.00      0.00        41
 NON-CRIMINAL (SUBJECT SPECIFIED)       0.00      0.00      0.00         2
                   NON - CRIMINAL       0.00      0.00      0.00        10
                        RITUALISM       0.00      0.00      0.00         6

                        micro avg       0.43      0.43      0.43   1559286
                        macro avg       0.22      0.13      0.13   1559286
                     weighted avg       0.38      0.43      0.37   1559286
```

The results seem a bit disappointing. Accuracy score barely increased from 43.47% to 43.49% compared to the first model. Classification report also shows very little improvement. Maybe in a large dataset like this, normal random forest classification already shows a good result.

c. K-Nearest Neighbors

The third and last model I tested is a knn model. I tried different n_neighbors value (3, 5 and 7) but got similar results (accuracy score at around 33%, 37%, 38% respectively).

The following results are accuracy_score, confusion_matrix and classification_report for n_neighbors = 5:

```
0.3714302571818127
[[42304  2076   393 ...     0     0     0]
 [ 3955 13420   489 ...     0     0     0]
 [ 1387   905   380 ...     0     0     0]
 ...
 [    0     0     0 ...     0     0     0]
 [    1     0     0 ...     0     0     0]
 [    1     0     0 ...     0     0     0]]
C:\Users\Guo\AppData\Local\Programs\Python\Python37\lib\site-packages\sklearn\metrics\c
1143: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.
no predicted samples.
  'precision', 'predicted', average, warn_for)
                                   precision    recall  f1-score   support

                  CRIMINAL DAMAGE       0.19      0.24      0.21    177212
                CRIMINAL TRESPASS       0.30      0.30      0.30     44774
                WEAPONS VIOLATION       0.06      0.02      0.03     16950
                          BATTERY       0.41      0.55      0.47    283261
                            THEFT       0.43      0.53      0.48    329930
                          ROBBERY       0.19      0.07      0.11     58770
               DECEPTIVE PRACTICE       0.32      0.19      0.24     62564
                          ASSAULT       0.12      0.04      0.06     96943
                        NARCOTICS       0.59      0.76      0.66    162251
              MOTOR VEHICLE THEFT       0.21      0.17      0.19     71204
                    OTHER OFFENSE       0.18      0.11      0.14     97006
                         BURGLARY       0.33      0.28      0.30     90178
 INTERFERENCE WITH PUBLIC OFFICER       0.00      0.00      0.00      3752
            PUBLIC PEACE VIOLATION       0.13      0.03      0.05     11223
        OFFENSE INVOLVING CHILDREN       0.10      0.01      0.02     11064
              LIQUOR LAW VIOLATION       0.28      0.05      0.08      3086
                        OBSCENITY       0.00      0.00      0.00       153
CONCEALED CARRY LICENSE VIOLATION       0.00      0.00      0.00        96
                         STALKING       0.00      0.00      0.00       809
                       KIDNAPPING       0.00      0.00      0.00      1375
                         HOMICIDE       0.95      0.28      0.43      2294
              CRIM SEXUAL ASSAULT       0.09      0.00      0.01      6526
                     PROSTITUTION       0.43      0.09      0.14     15142
                      SEX OFFENSE       0.15      0.00      0.00      5842
                            ARSON       0.06      0.00      0.00      2553
                     INTIMIDATION       0.00      0.00      0.00       877
                 HUMAN TRAFFICKING       0.00      0.00      0.00        14
                         GAMBLING       0.08      0.01      0.01      3304
                     NON-CRIMINAL       0.00      0.00      0.00        42
          OTHER NARCOTIC VIOLATION       0.00      0.00      0.00        32
                  PUBLIC INDECENCY       0.00      0.00      0.00        41
 NON-CRIMINAL (SUBJECT SPECIFIED)       0.00      0.00      0.00         2
                    NON - CRIMINAL       0.00      0.00      0.00        10
                        RITUALISM       0.00      0.00      0.00         6

                        micro avg       0.37      0.37      0.37   1559286
                        macro avg       0.16      0.11      0.12   1559286
                     weighted avg       0.33      0.37      0.34   1559286
```

As the results show, accuracy score is not as good as previous models and classification_report scores are worse in general as well. However, classification_report showed similar pattern as previous models and the precision score for “Homicide” is actually the highest among the three models. In addition, this knn model used least amount of time to complete while the second model took the longest time. So the knn model is actually the more time efficient one for this analysis. I also noticed much less RAM use while running the knn models.