Project Report

IMT 574

Weining Xu

Abstract

The project is about Chicago crime prediction. We want to predict the crime types when knowing the location of the crime happens in the city of Chicago, so that it could help the police dealing with the coming crimes by acknowledging the possible types of crime might be, and taking action more effectively, like having fire fighters or ambulance ready. Based on the prediction, the government can also make decisions on increasing police force in some certain area in a seasonal pattern and having more police portal at certain time of each day. Hoping this could help reduce the crimes and provide a safer community in Chicago.

Data Collection

We download the data from Chicago data portal [1], which has 7.29 million rolls and 22 columns each role is a reported crime occurred in the city of Chicago from 2001 to Feb 24th, 2021. Each column represents: 'ID', 'Case Number', 'Date', 'Block', 'IUCR', 'Primary Type',

'Description', 'Location Description', 'Arrest', 'Domestic', 'Beat',

'District', 'Ward', 'Community Area', 'FBI Code', 'X Coordinate',

'Y Coordinate', 'Year', 'Updated On', 'Latitude', 'Longitude', 'Location'

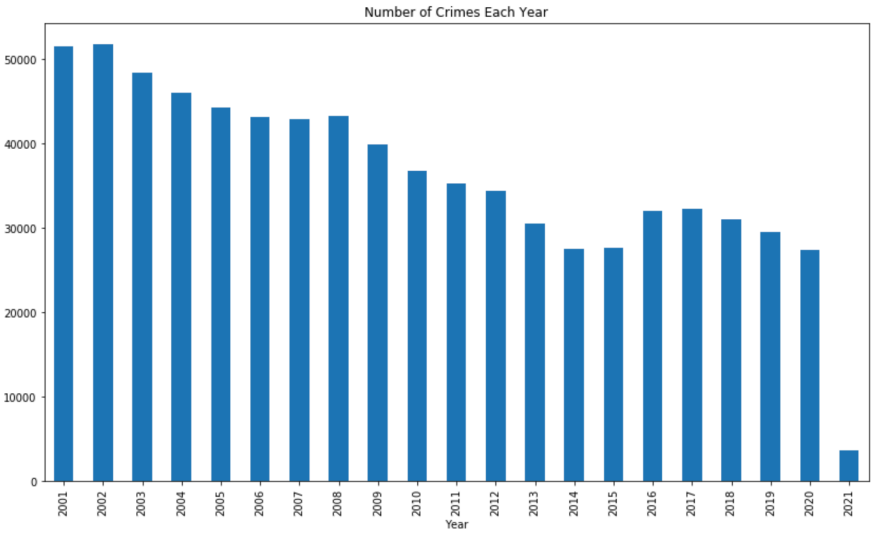
Since we want to predict crime types based on locations, we will keep all features relevant to locations like 'Arrest', 'Beat', 'District', 'Ward', 'Community Area', 'X Coordinate',

'Y Coordinate', 'Latitude', 'Longitude', 'Location'.

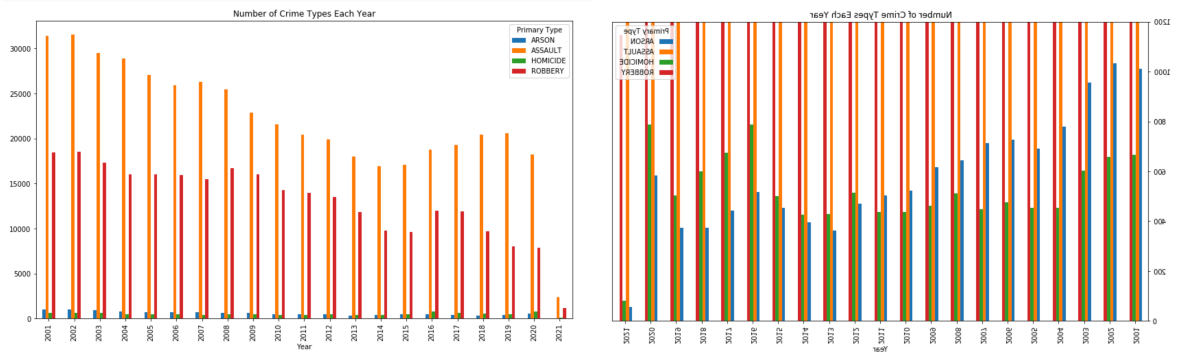
We also keep 'Date', 'Year', 'IUCR', 'Primary Type' for time and crime identification.

Data Visualization

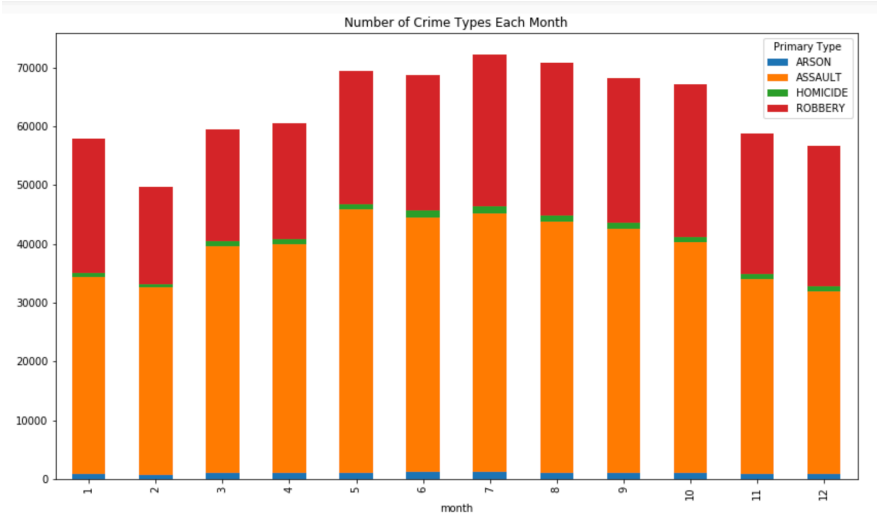
Before doing the prediction, let's see the total number of crimes each year which we can see that it is decreasing over these years which is good for the society, but we want it keep decreasing. There are many crime types in the data, but we only focus on four types of crimes they are homicide, assault, robbery and arson, we labeled them from 1 to 4.

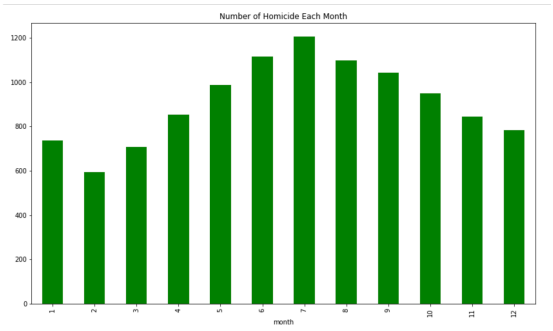


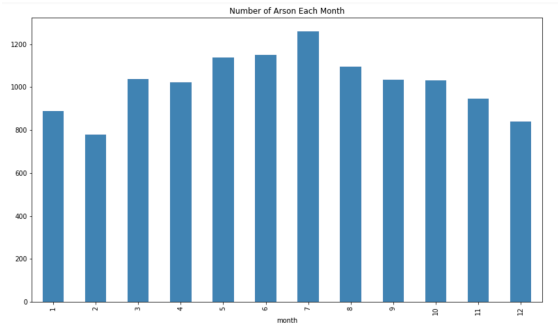
Then the following left plot describe the amount of 4 types of crimes happened each year. We can observe that assault and robbery are the most occurred incidents among four types. These two are having the same decreasing trend as the previous plot. Then let's zoom in, on the right, we can see arson is also having a decreasing trend, but homicide is increasing clearly in the past five years.

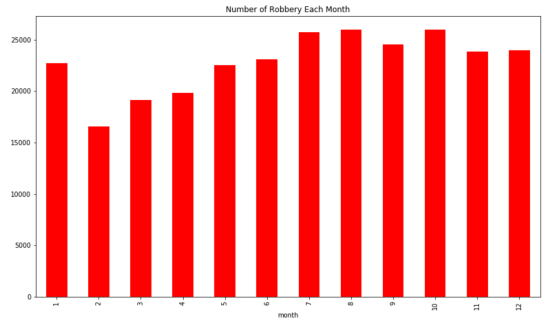


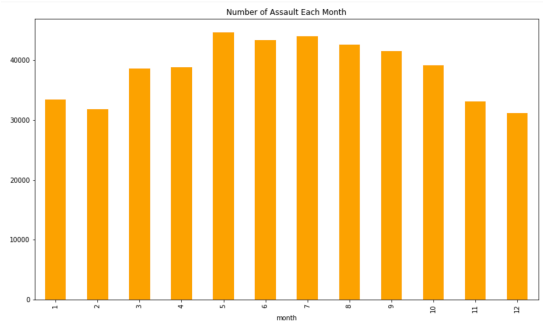
We also want to figure out if there's a seasonal pattern of each crime types occurred. Overall, we can see that from May to October which is summer and fall there are more crimes occurred compared to winter and spring. So as homicide and arson, they are having a clear pattern of occurrence in July, robbery has a drop in spring, and assault has a slight drop in winter. Also, each has an obvious drop in February.











After seeing the pattern of each types of crimes we can start finding the features for prediction since we know our target is the types of the crimes. Before choosing the features, let's see the correlation plot of each relevant features first. I have year month day extracted from date. We can see that features of arrest, year, month and day are barely correlated with other features, so we drop them. Also, x-coordinate and the longitude have correlation equals to one, so do y- coordinate and latitude. We only keep XY coordinates. Therefore, the final features we are using for prediction are: 'Beat', 'District', 'Ward', 'Community Area', 'X Coordinate', 'Y Coordinate'.

	Arrest	Beat	District	Ward	Community Area	X Coordinate	Y Coordinate	Year	Latitude	Longitude	month	day
Arrest	1	0.018	0.016	0.023	-0.018	-0.0051	0.023	-0.038	0.023	-0.0047	-0.016	-0.00086
Beat	0.018	1	0.94	0.66	-0.5	-0.53	0.63	-0.042	0.63	-0.53	1.3e-05	0.00038
District	0.016	0.94	1	0.7	-0.49	-0.58	0.63	-0.009	0.63	-0.58	0.00019	0.00052
Ward	0.023	0.66	0.7	1	-0.52	-0.5	0.62	0.015	0.62	-0.5	-2.2e-05	-5.1e-05
Community Area	-0.018	-0.5	-0.49	-0.52	1	0.32	-0.77	-0.017	-0.77	0.31	-0.0034	0.00051
X Coordinate	-0.0051	-0.53	-0.58	-0.5	0.32	1	-0.48	-0.0052	-0.48	1	0.0013	0.0014
Y Coordinate	0.023	0.63	0.63	0.62	-0.77	-0.48	1	0.001	1	-0.47	0.00043	-0.00018
Year	-0.038	-0.042	-0.009	0.015	-0.017	-0.0052	0.001	1	0.001	-0.0053	-0.014	0.00095
Latitude	0.023	0.63	0.63	0.62	-0.77	-0.48	1	0.001	1	-0.47	0.00043	-0.00018
Longitude	-0.0047	-0.53	-0.58	-0.5	0.31	1	-0.47	-0.0053	-0.47	1	0.0013	0.0014
month	-0.016	1.3e-05	0.00019	-2.2e-05	-0.0034	0.0013	0.00043	-0.014	0.00043	0.0013	1	-0.0041
day	-0.00086	0.00038	0.00052	-5.1e-05	0.00051	0.0014	-0.00018	0.00095	-0.00018	0.0014	-0.0041	1

Data Modeling

Since the target is categorical data, we want to use supervised learning techniques with classification. We will apply naïve bayes, KNN, and random forest as algorithms to predict the types of crimes in the last week of 2021.

Naïve Bayes

Naïve bayes is a fast classification algorithm for binary (two-class) and multi-class classification problems, which is useful in this situation for we want to classify crime types.

Multinomial Naïve Bayes

First, we use Multinomial Naïve Bayes algorithms to build train the data, for the target is discrete numbers from 1 to 4. We only get 0.02 accuracy score which is super low, since our features contains continuous data, which Multinomial Naïve Bayes is not a good fit for predicting crime types.

```
                precision    recall  f1-score   support

             1       0.02      0.67      0.03         6
             2       1.00      0.01      0.01       375
             3       0.00      0.00      0.00       138
             4       0.02      0.50      0.04        10

      accuracy                           0.02       529
     macro avg       0.26      0.29      0.02       529
  weighted avg       0.71      0.02      0.01       529

[[  4   0   0   2]
 [185   2   8 180]
 [ 64   0   0  74]
 [  5   0   0   5]]
```

Gaussian Naïve Bayes

Then we use Gaussian Naïve Bayes algorithms to fit the data. We get 0.71 accuracy score which is high. But when we look at the confusion matrix, we can see that it just predicting every crime to be

assault, which is based on the large portion of assault crimes in training and testing set. But it is more like guessing so I don't feel like it is in appropriate approach.

```
                precision    recall  f1-score   support

             1       0.00      0.00      0.00         6
             2       0.71      1.00      0.83       375
             3       0.00      0.00      0.00       138
             4       0.00      0.00      0.00        10

      accuracy                           0.71       529
     macro avg       0.18      0.25      0.21       529
  weighted avg       0.50      0.71      0.59       529

[[  0   6   0   0]
 [  0 375   0   0]
 [  0 138   0   0]
 [  0  10   0   0]]
```

Categorical Naïve Bayes

Then I tried Categorical Naïve Bayes [2], which is suitable for classification with discrete features that are categorically distributed. We use them since most features are categorized. It has 0 .65 accuracy score. When we look into the confusion matrix it has all six homicide crimes and 10 arson crimes predicted wrong. It has a high accuracy score only because of it has most of the assault crimes correct.

```
                precision    recall  f1-score   support

             1       0.00      0.00      0.00         6
             2       0.74      0.83      0.78       375
             3       0.36      0.28      0.31       138
             4       0.00      0.00      0.00        10

      accuracy                           0.66       529
     macro avg       0.27      0.28      0.27       529
  weighted avg       0.62      0.66      0.63       529

[[  0   5   1   0]
 [  0 311  64   0]
 [  0 100  38   0]
 [  0   6   4   0]]
```

KNN Classification

Same situation happens in KNN classification with number of neighbors equals 4. The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. KNN captures the idea of similarity (sometimes called distance, proximity, or closeness) with some mathematics calculating the distance between points in order to classify them into groups. [3] The accuracy score is 0.68 but it has all six homicide crimes and 10 arson crimes predicted wrong. Probably because most of the crimes in the training set are assault and robbery so the algorithm is learning these two crimes better.

```
              precision    recall  f1-score   support

           1       0.00      0.00      0.00         6
           2       0.74      0.85      0.79       375
           3       0.39      0.27      0.32       138
           4       0.00      0.00      0.00        10

    accuracy                           0.67       529
   macro avg       0.28      0.28      0.28       529
weighted avg       0.63      0.67      0.64       529

[[  0   3   3   0]
 [  5 317  52   1]
 [  0 101  37   0]
 [  0   7   3   0]]
```

Random Forest

Then I apply random forest algorithms to make the prediction. Random forest operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes for classification task or mean prediction of the individual trees when used for regression tasks. We have 0.64 accuracy score when we iterate it 30 times with 100 estimators. From the confusion matrix although it got all arson crimes wrong, but it got 17% of homicide crimes correct and 76% of the assault crimes correct and 43% of the robbery crimes correct. Even though the accuracy score is not the highest, but it has a good result in categorizing three types of crimes.

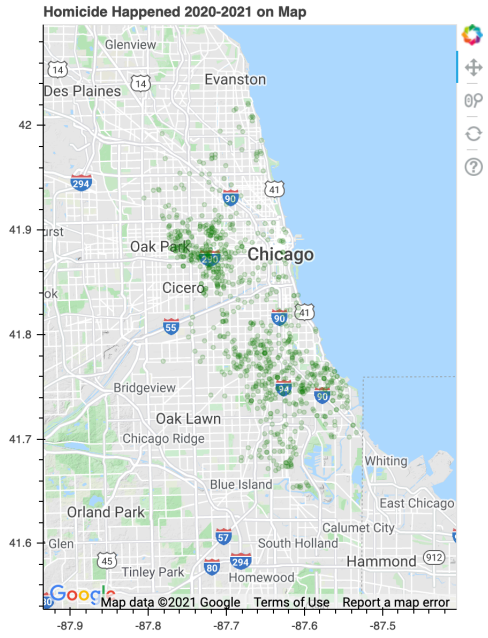
```
              precision    recall  f1-score   support

           1       0.12      0.17      0.14         6
           2       0.77      0.76      0.76       375
           3       0.40      0.43      0.42       138
           4       0.00      0.00      0.00        10

    accuracy                           0.65       529
   macro avg       0.32      0.34      0.33       529
weighted avg       0.65      0.65      0.65       529

[[  1   2   3   0]
 [  7 285  82   1]
 [  0  78  60   0]
 [  0   6   4   0]]
```

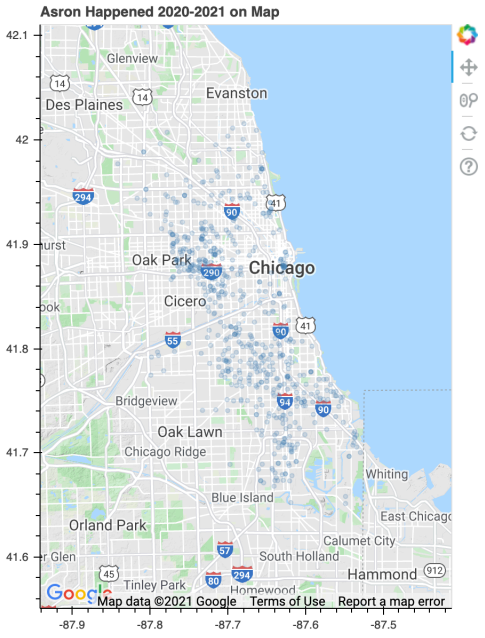
Conclusion

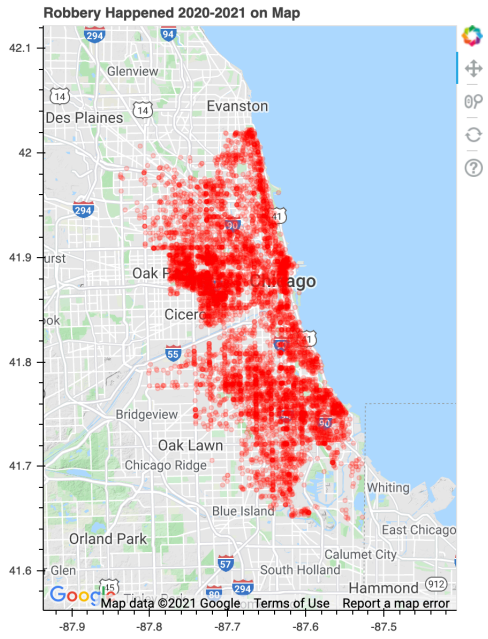
Therefore, based on their accuracy score we can say Gaussian naïve bayes model is the best but based on Confucian matrix I think random forest model is the best.

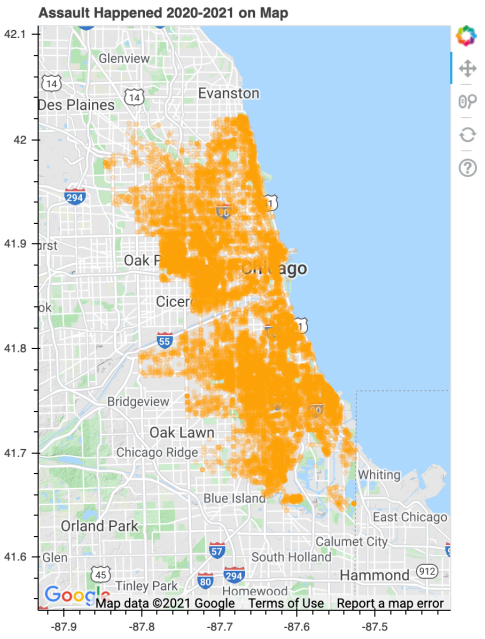
Location Visualization

At last, I plot each type of crimes happens from 2020 till now on the map. [4] We can see many homicide and arson crimes happened more focused around Oak Park.









Improvement

If we want to look for further improvement in the prediction of crime types, we can find housing price in each area, poverty rate and unemployment rate in each area and map them into the data set and hope it can help figuring out the improvement of the accuracy score for predicting crime types.

Source:

[1] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

[2] https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.CategoricalNB.html#sklearn.naive_bayes.CategoricalNB

[3] https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[4] https://thedatafrog.com/en/articles/show-data-google-map-python/