Name: Adigun Daniel Oluwasegun

ECON9000 FINALS

Data Description

The raw data was gotten from https://we.tl/t-9q2e1SRVpR), this dataset represent crime data for the city of Chicago from 2001-to date(with the exemption of sensitive information especially for active cases).
The data set had millions of entry so the first thing I did was write a code (crimes_data_cleaning.py) that broke the data down to years I was interested in and began the cleaning process.
The first part of the cleaning was restricting my data set to crimes recorded between 2017-2018 and use that for my predictive analysis, next step was to take care of all missing values and ensure all the datasets could be run of the various packages that there were going to be tested on.
From a dataset with about 30 labels, I restricted the dataset to contain the following labels (Arrest, Domestic, District, Ward, Community Area, FBI Code, Census Tracts, and Police Districts)
These labels were chosen for the following reasons

i. the FBI Code Identified the crime committed so we had no use for any other labels describing the same incident. There are a total of 26 codes for these identifying various crimes at various locations, Table 1. below shows a breakdown of each code and crime it identifies.

ii. The Ward label identified the location of each of the crime incidents recorded. The city of Chicago has a total of 50 wards with each number identifying a specific ward.

iii. The Arrest and Domestic Labels were texts (True or False) which were recorded to 1 or 2, so we don't have any issues running the data through any of our models.

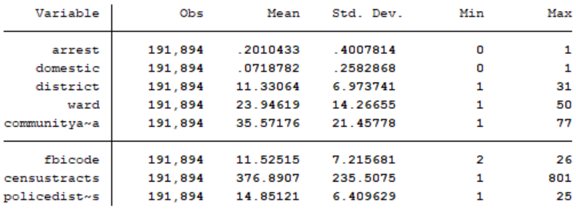
iv. The census tracts was chosen, under the assumption that arrears with higher populations will have higher crime rates values between 1-800 were recorded under this label

v. The police districts identified the police sector of the state that handled the case and patrols a specific ward or community area

vi. Community area was an additional location Identifier with the city of Chicago having a total of 77 community areas.

FBI Code	Crime Description
02	Criminal sexual assault
03	Robbery
05	Bulglary
06	Larceny
07	Motor vehicle theft
09	Arson
10	Forgery and counterfeiting
11	Fraud
12	Embezzlement
13	Stolen property
14	Vandalism
15	Weapons violation
16	Prostitution
17	Criminal sexual abuse
18	Drug abuse/ narcotics
19	Gambling
20	Offense against family
22	Liquor license
24	Disorderly conduct
26	Misc non-index offense

Table 1: FBI Code for criminal offense



```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      arrest |    191,894    .2010433    .4007814          0          1
    domestic |    191,894    .0718782    .2582868          0          1
    district |    191,894    11.33064    6.973741          1         31
        ward |    191,894    23.94619    14.26655          1         50
communitya~a |    191,894    35.57176    21.45778          1         77
-------------+---------------------------------------------------------
     fbicode |    191,894    11.52515    7.215681          2         26
censustracts |    191,894    376.8907    235.5075          1        801
policedist~s |    191,894    14.85121    6.409629          1         25
```

Figure 2: Descriptive statistics of variables

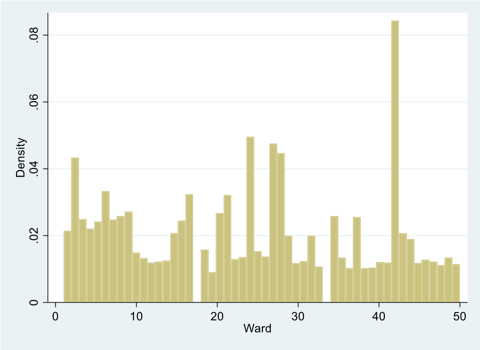


Figure 2: histogram showing wards and frequency of reported crime incidents between 2017-2018

ANALYSIS

The variables/labels collected and cleaned were chosen for specific reasons to aid our analysis, first its been long debated that increased policing could potentially increase the crime rate in cities, this is a well-studied topic under econometrics, hence our inclusion of police districts in the dataset to see how it affects the crime rate in certain wards, also with the FBI Code we will be able to identify specific crimes and possibly tie them to areas where the occurred frequently. Our final assumption is that higher populations could potentially increase the crime rates in certain wards/communities hence our inclusion of this variable in the analysis.

This dataset will be analysis using the following models:

1. Linear regression.
2. Random Forest.
3. Decision tree.
4. KneighborsClassifier.

Each of these models were compared using certain parameters observed after running the respective models. Mean squared Error, R-squared, Accuracy Score, and confusion matrix to see how often predictions are correct.

We also performed some tuning for some of the models in order to have the best parameters obtainable for each of the models, this process was carried out for the knn_model with the k_neighbors values tested from 5-8 and the highest accuracy score obtained for k_neighbor value of 7.

K_Neighbor Value	5	6	7	8
MSE	6.911703	6.821798	6.794957	6.780949
Accuracy	0.815087	0.815504	0.819381	0.817484
R2_score	0.764465	0.770553	0.772355	0.773292

Table 2: comparative results of KNN tuning

A similar approach was done for the Random_forest n_estimator, we varied its value between 11 and 14 and we found that the best parameters were obtained at the n_estimator value of 12, below is a breakdown of the changes in parameters when this was done.

N_estimator (Random Forest)	11	12	13	14
MSE	6.359978	6.350872	6.389598	6.373805
Accuracy	0.835014	0.835098	0.834577	0.834410
R2_score	0.801151	0.801720	0.799295	0.800286

Table 3: comparative results of Random Forest Tuning

Evaluating Model	Linear Regression	KNN	Decision Tree	Random Forest
MSE	11.112631	6.794957	6.374145	6.350872
Accuracy	-	0.819381	0.833680	0.835098
R2_score	0.391139	0.772355	0.799678	0.801720

Table 4: Comparative result of Machine Learning models and their performance.

The comparative result shows that the Random forest model gave the best results In terms of accuracy, r2_score and MSE, with the linear model giving the least convincing results and can be identified as the least reliable model of all the four models used for our analysis. It is also important to note that random forest takes multiple trees into account and gives an average result which proved to be perfect for this data.