Sai Trisha Pusuluri

11373822

**CRIME PREDICTION IN CHICAGO**

Introduction:

The quality of life of the individuals has been unfavorably influenced by the crimes in the networks. The quest for bliss and fulfillment is contaminated by the dread and vulnerability of things to come among individuals. Numerous legislative organizations have been working enthusiastically to give the suspicion that all is well and good and to reestablish the confidence locally and the framework. Alongside the help and collaboration from the general population, these offices require a ton of monetary and specialized assets to get proactive and rout the crooks. The assortment of past criminal records and their investigation can give specialized favorable position to the legislative offices. Finding the examples in violations, making forecasts on conceivable future wrongdoings dependent on those examples can be done by utilizing different machine learning.

statement of the problem:

The statement of the project problem is that I will decide the police beat of a specific crime relying upon variables, for example, essential sort of crime, crime depiction, area, ward, local area, and a gathering of different highlights which will be portrayed in forthcoming areas.

Literature Review:

1. There have been various examination works including data analysis on the crime and also the prediction of crime. Some of them are near what we have attempted to execute in this paper of Almanie et al. 2015; Luna 2019 depending on the data type and tackling the problem. Whereas others Sathyadevan et al. 2014; Bogomolov et al. 2014. That have attempted to zero in on comparative issues utilizing various organizations of data and various methods of learning when contrasted with our usage. In this segment, we will examine the short synopsis of the techniques that these works have executed.

2. From the Almanie et al. 2015 have attempted to misuse the spatial and worldly highlights in crime datasets to improve their model&#39;s prescient capacity. They have completed measurable correlation of two distinctive crime datasets (Denver Crime Dataset, and Los Angeles Crime Dataset) and utilized from the earlier algorithm to get the designs in crime. There are 333k rows and The Denver Crime Dataset comprised of 333k occurrences with attributes (19) and the LA dataset comprised of 243k examples with 19 highlights. These numbers are altogether more modest than the dataset we utilized (roughly 7 million cases). Taking a gander at the span of information assortment, Denver has information from the time frame years (2010-2015) and 96 percent of the examples in LA dataset were gathered from 2014 through 4 percent were gathered before 2014. Logistic regression is the additionally utilized different AI techniques, XGBoost and Random backwoods classifier while the later playing out the best with 72% F1-score.

Objective of the study:

The main objectives of my work are:

1.To estimate the crimes rate dependent on their frequencies over years with knowledge on how the number crimes fluctuates over years and over seasons.

2. Through the various geographical areas namely, locality, area, district the crime frequencies are analyzed.

3. To understand the examples of major crimes including arrest over years with execution of strategies to foresee the crimes including capture.

The initial step is getting a decent understanding on data analytics of crime, a comprehension of what factors by and large are acceptable indicators of crime area. We will at that point design the information as we would prefer by cleaning the information, and filling in missing qualities as we see vital, this will prompt element designing and permit us to comprehend and distinguish designs in the dataset. Ultimately, we will utilize algorithms like Decision tree and logistic regression for anticipating the beats certain crimes will occur. The problem here is exciting and challenging, thus three various objectives are estimated.

Dataset used:

The data set is collected from Kaggle which is continuous update and data source is city of Chicago. It has the rows of 6.99m rows and has 22 columns. The columns are.

Date- Date nothing but the date of incident happened.

Block- where the crime happened in the address block.

Iucr- Illinois Uniform Crime Reporting

Description- IUCR code secondary description, Criminal sexual maltreatment, provocation by electronic methods, discovered presume opiates, and so on

Primary type- IUCR code primary description, sex and various offences.

Arrest- It shows the arrest is done or not

Beat- police patrolling.

Year- The year of crime happened.

Ward, x location, y location, location, updated record

From the data set there are 20 years,12 months, 4 seasons, 23 districts, 7 types of primary grouped, location, x coordinate, y coordinate, 370 iucr, 34 primary type (includes all types of crimes), 450 description, 2 types of arrest, too many and continuous updated on.

There are many locations in the dataset with continuous variable, too many crimes per year.

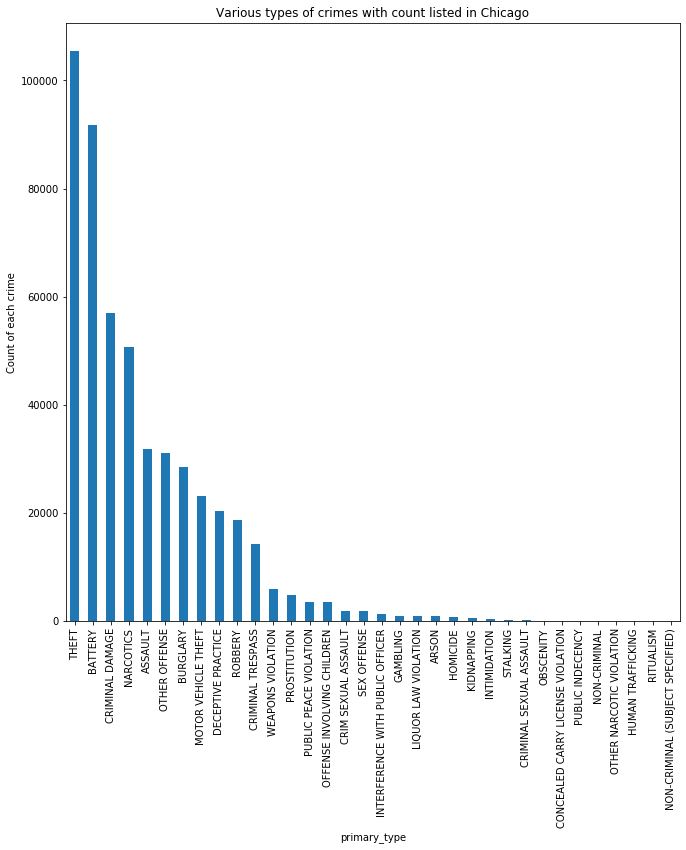
The underlying dataset contains around 7 million records, yet this requires very nearly 20-25 minutes to run for each cycle. So, we chose to test information to 500,000 columns additionally this is peer input that we got. Rather than attempting to anticipate the beat, we are presently attempting to foresee whether a culprit will be captured for the episode they submit dependent on the highlights above.

We divided the information as 70 % for Training information, 20% for tuning the information and the remainder of the 10% for testing the information. To keep up free and indistinguishably dispersed information split we get truly near a 70/20/10 split. The preparation dataset will have information cases that happen inside the years 2001-2012, the tuning dataset will contain information from 2013-2017, and finally the testing dataset will contain information from 2018-2020.

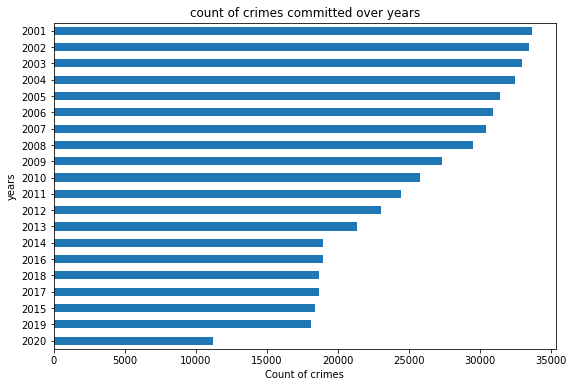
Exploratory data analysis (EDA):

The data set contains 22 columns and 6.99m rows. In the data cleaning, I have removed the unwanted rows and columns all the null values are removed. Here I separated the month, year and then according to the month I have added a new column season. After the data cleaning there are 10 columns and 500000 rows.

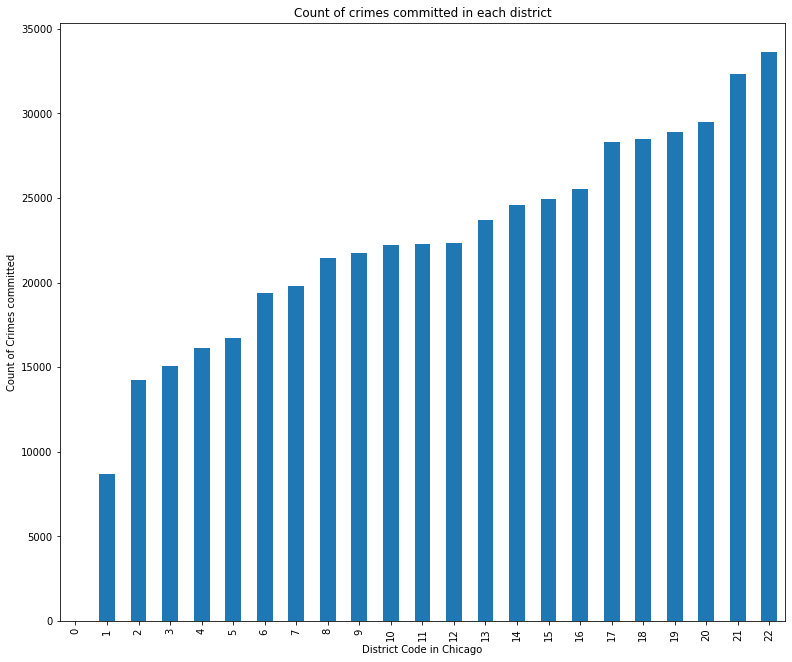
Hypotheses for the Study:



The above bar graph illustrates the different types of crimes with count listed in Chicago from the year 2001 to 2020. The x-axis indicates the primary types of the crimes whereas the y-axis indicates the count of each crime. Theft is the crime, which is recorded with the highest count of 110000, second is the battery with the count of 9lakh and then the criminal damage is the next most frequently committed crime with the count of approximately 60000. The least committed crime in Chicago is obscenity, concealed carry license violation, public indecency, non-criminal, other narcotic violation, human trafficking, ritualism.

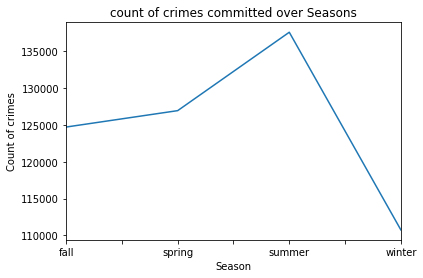


The above bar graph represents the count of crimes committed over years. The x-axis represents the count of the crime, the y-axis indicates years. In the year 2001 crime rate is excessive with the count of 34000 whereas in the 2002 the count of the crimes is slightly reduced to 32000. The crime counts are gradually decreasing year by year. From the year 2014 to 2019 the count of crimes are stable with around 18000. The count dropped down to 11000 in the year 2020.

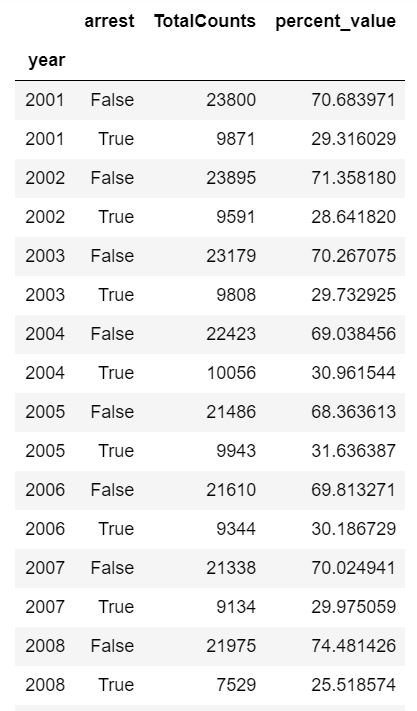


The Bar graph illustrates the count of crimes committed in each district. The The x-axis represents the Chicago district code, the y-axis represents committed crimes count. The crimes are high in the district 22 with the count of 320000 crimes and there are least crimes in the district 1 with 90000 crimes.

Data Analytics:

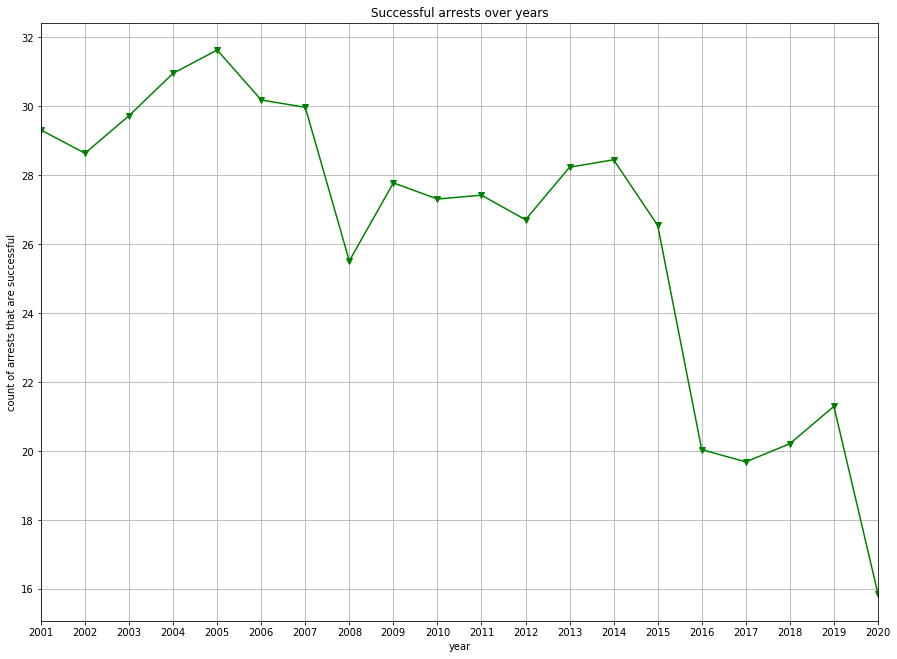


The line graph illustrates the count of crimes committed over seasons. The x-axis represents the seasons and the y-axis represents the count of crimes. In the fall season the crimes are 125000, in the spring season it raised little high to 127000 whereas in the summer season it drastically raised to 140000 and then the crime is very low in the winter season.

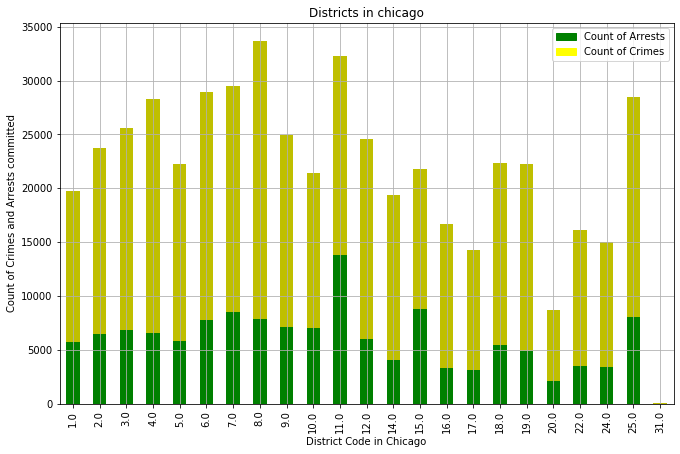


From the table seen above tells the criminals arrested and not arrested in the years 2001 to 2020 along with the percentage value. In the year 2001, 23800 people were not arrested and then 9871 people are arrested. Correspondingly, in 2002 to 2020 the left-over people are more same as the year 2001. The crimes are decreasing as the year increases. In the year 2020, the 9426 people were left not arrested whereas 1778 people are arrested.

Data Visualization:



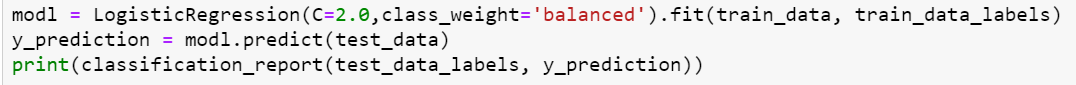
The line graph represents the successful arrests over the years. The x-axis indicates the years, and the y-axis indicates the count of arrests that are successful. In the year 2005 there are more arrests compared to other years. There are a smaller number of arrests in the year 2020.



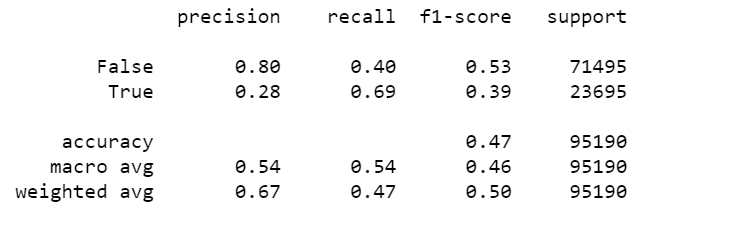
This tells the districts in Chicago with the number of arrests and the crimes. The districts code is considered in the x-axis and the number of arrests and the crimes are considered as the y-axis. More crimes are committed in the district 8.0 with 34000 but less criminals are arrested with around 9000. In the district 11.0 more number of crimes are committed and more criminals got arrested. In the district 20.0 less crimes are committed and less criminals are arrested.

Results report:

Logistic regression is nothing but the algorithm classification which helps to analyze the likelihood of a reliant variable. The reliant variable is a binary variable has the coded data. For this regression we need to collect all the data required, then data should be imported into the python by using the pandas, sklearn, seaborn. Then logistic regression must be created.



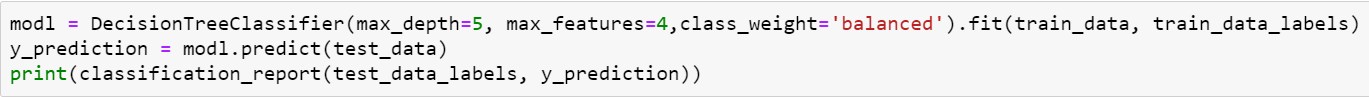
Results:



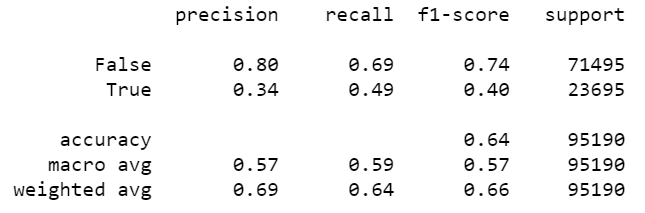
Here I have only considered the arrest for all the years (2001 to 2020). From above we can say that the accuracy is less with 47%. The macro average is 54% , weighted average is 67. I can also say that there are 71495 not arrested criminals in the year 2001 to 2020 whereas 23695 arrested in the same years.

Decision Tree:

They are non-parametric managed learning technique utilized for arrangement and relapse. The objective is to make a model that predicts the worth of an objective variable by taking in basic choice principles surmised from the information highlights.



Results:



From above we can say that the accuracy is less with 64%. The macro average is 57% , weighted average is 66. I can also say that there are 71495 not arrested criminals in the year 2001 to 2020 whereas 23695 arrested in the same years.

Conclusion:

I can conclude by saying that I have attempted to show why there is a tremendous disparity in the capture class circulation, From the data set I have shown some degree indistinguishably appropriated. How the crime changed year to year and based on seasons. Ultimately, With the help of Chicago crime data we have anticipated the numbers of criminals were arrested and the number of criminals were left freely. From the data set the columns are increased by splitting the month, year, time, and the season column is added to the data set based on the months. We have considered the decision tree classifier and logistic regression. Finally, I would like to say that the accuracy is higher in the decision tree with 64% and the accuracy in the logistic regression is less with 47%. So, the decision tree is most preferrable for the crime prediction than the logistic regression.

Bibliography:

* (<https://www.kaggle.com/chicago/chicago-crime>) Dataset
* https://www.irjet.net/archives/V5/i9/IRJET-V5I9192.pdf
* <https://www.researchgate.net/publication/335854157_Predicting_Crime_Using_Time_and_Location_Data>
* https://www.wired.com/insights/2013/08/predictive-policing-using-machine-learning-to-detect-patterns-of-crime/