Chicago Crimes

Data Source

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

Objective:

This dataset includes reported incidents of crime (with the exception of murders, for which data is available for each victim) that occurred in the City of Chicago between 2001 and the present, minus the most recent seven days. The information is derived from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. Addresses are shown at the block level only to protect the privacy of crime victims, and specific locations are not identified.

Data Description:

1. ID - Unique identifier for the record.

2. Case Number - The Chicago Police Department RD Number

3. Date - Date when the incident occurred. this is sometimes the best estimate.

4. Block - The partially redacted address where the incident occurred.

5. IUCR - The Illinois Uniform Crime Reporting code.

6. Primary Type - The primary description of the IUCR code.

7. Description - The secondary description of the IUCR code, a subcategory of the primary description.

8. Location Description - Description of the location where the incident occurred.

9. Arrest - Indicates whether an arrest was made.

10. Domestic - Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.

11. Beat - Indicates the beat where the incident occurred.

12. District - Indicates the police district where the incident occurred.

13. Ward - The ward (City Council district) where the incident occurred.

14. Community Area - Indicates the community area where the incident occurred.

15. FBI Code - Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).

16. X Coordinate - The x coordinate of the location where the incident occurred

17. Y Coordinate - The y coordinate of the location where the incident occurred

18. Year - Year the incident occurred.

19. Updated On - Date and time the record was last updated.

20. Latitude - The latitude of the location where the incident occurred

21. Longitude - The longitude of the location where the incident occurred.

22. The location where the incident occurred in a format that allows for the creation of maps and other geographic operations on this data portal.

Exploratory Data Analysis

Using this data frame, I have shown the basic visualizations. We have used different types of visualizations where the data can be described in different ways. It shows the basic data exploration and many statements. And here while I was doing the visualizations, I faced problems due to a large number of categorical columns and values. While doing some plots it was showing the data as bulk and cannot get the proper visualizations. So, in this case, I need to take the value count to a limit from the columns that I need to visualize.

Modeling:

I choose "Arrest" as my target variable for modeling. And I have created pipelines for modeling but we don’t have any numerical features in the dataset so, we have only categorical variables we created a categorical pipeline.

Feature selection for the modeling:

Since we are restricting the problem data to that which can be obtained , we are gonna drop all but those variables. As for the target variable, we are only considering the 'Primary Type'.

We have used the below columns as they have more correlation with the primary type ‘Arrest’

Case Number object

IUCR object

Primary Type object

Arrest bool

Domestic bool

FBI Code object

Model Selection:

As we observe here "logistic regression model" performs well as it got 90% accuracy. So, I selected this as the best model for my data

Summary:

1. Logistic regression performs well

2. The roc accuracy score is higher for the logistic regression as it maintains 90.01%

3. Coming to the crimes most of the crimes are theft based, in the 10th district the crimes are high, and also most of the crimes occurred in the 10th month i.e., October.

Results:

Logistic Regression Classifier

1. Logistic Regression Score: 88.59%

2. ROC-Accuracy Score: 90.10%

3. F1-Score: 72.61%

Further Exploration and Features:

1. I want to explore more visualizations in the final submission.

2. As after observing the data set while doing the visualizations I want to consider the "Description" columns as my target variable to get the code to be explored.

3. Here I have no chance to use the column "Location Description" as we have a lot of null values and no chance to get the perfect visualizations.

4. In the future project I am going to analyze the accuracy of the crimes using the Description as the target variable, and also going to use the particular classifiers and machine learning problem curve depending upon the data.

5. I am going to split the data and create a pipeline and also find out the accuracy of test and train data sets.

6. On a note I just want to let you know that I could generate more visualizations and process the data ways while doing the final project.

7. A typical approach is, to improve a model to use more data. In this case, I have only taken the chunk data i.e., up to 100000 but including the whole can improve the overall score as taking the chunk lot of data got missed like a number of years or months.

8. We could try various techniques, such as a complicated neural network, but this could lead to additional problems and is overkill.

9. For more exploration we can use ensemble models, by using this the scores may get improved.

References:

1. https://www.geeksforgeeks.org/how-to-load-a-massive-file-as-small-chunks-in-pandas/

2. https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LogisticRegression.html

3. https://scikit-learn.org/stable/modules/tree.html

4. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

5. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc\_curve.html