Crime Prediction

Project URL:

https://github.com/jwchoi3120/Big-Data-Projects/tree/main/cse482-project

ABSTRACT

In this project, our team have decided to predict the crime rate for 2020. We have collected full data for 2018, 2019 and part of 2020. Our goal was to practice using sklearn library with all the prediction with using training set and test set. We could have achieve our goal which we got way more familiar with prediction model and we got the result which we could definitely see that can be happening in 2020.

1. INTRODUCTION

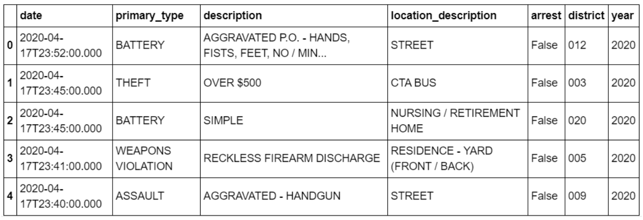
1. Chicago is one of the biggest city in United States. There are so many crimes going on, and our team wanted to predict what is going to happen after the time that we collected data. It is very important to predict those since there should be enough cops or support systems ready for those situations.
2. Our goal is to use the data from Chicago's API to predict crime rates. We have used data from the city of Chicago using their API to get the data. We have decided to use the arrest data, the date of the crime, the description of the crime. We think these descriptions will be most useful in predicting crime rates.
3. We have downloaded the data using the API and put it into a dataframe. We divided the data into training, validation and testing sets. We have used a sklearn library and we cleaned the data as we trained the model.
4. We have extracted the data from the Chicago Police Department. We have extracted 600,000 cases from the file because the file was large, and it contained full crimes that happened from 2008 to 2020.
5. When we were reading the csv file from the Chicago Police Department, the file was too big to process. We have waited for a long time, and it still couldn't process it, so we have set our limitation to 600,000. We thought we had gotten enough data from that.
6. The predicted result was extremely similar to the actual data, so we would say it was very successful. We have achieved our goal which was predicting the crime rates for this year.

2. DATA

The data was extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. The dataset reflects reported incidents of crime that occurred in the City of Chicago from 2001 to present. There are more than 7 millions cases, but the size of the file is too big to analyze and develop a prediction model. For the analysis, the most recent 600,000 cases were selected. The time period of the dataset is from December 2017 to April 2020.

We were able to access the dataset via the Socrata Open Data API (SODA). The format of the data was JSON, and the Python package Pandas makes it easy to work with JSON data.

Our dataset has 23 features (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). There are some irrelevant features that must be discarded. We dropped all the columns, except for date, primary type, description, location description, arrest, district, and year.



	date	primary_type	description	location_description	arrest	district	year
0	2020-04-17T23:52:00.000	BATTERY	AGGRAVATED P.O. - HANDS, FISTS, FEET, NO / MIN...	STREET	False	012	2020
1	2020-04-17T23:45:00.000	THEFT	OVER $500	CTA BUS	False	003	2020
2	2020-04-17T23:45:00.000	BATTERY	SIMPLE	NURSING / RETIREMENT HOME	False	020	2020
3	2020-04-17T23:41:00.000	WEAPONS VIOLATION	RECKLESS FIREARM DISCHARGE	RESIDENCE - YARD (FRONT / BACK)	False	005	2020
4	2020-04-17T23:40:00.000	ASSAULT	AGGRAVATED - HANDGUN	STREET	False	009	2020

Table 1: Attributes of the data acquired from Chicago Data Portal

Besides irrelevant features, we faced other issues in the dataset. First, rows that contain null values should be removed because it may cause a possible problem when training the prediction model.

3. METHODOLOGY

We downloaded csv files for the data between 2008 and 2017 and used the Socrata API to download 2018 to 2020. We ran into problems when we tried to download more data from the API. It was too much to download. The data is the same set of data we just downloaded it in two different ways. We decided to analyze the primary type of battery because it had the most cases. We used linear regression of the previous twelve months for our prediction. We used the data from the csv files as training data because we had more data and we used the data from 2018 till now for the testing data. We used sklearn linear regression to solve this problem.

I created the X training and testing sets by making a table of the number of battery cases from the previous twelve months and used the values of the current months as the Y training and testing sets. The training set consisted of years 2008 to 2017 and the testing set consisted of years 2018 to 2020.

All of the code for this project is written in 482.project.ipynb. The data from 2008 to 2017 is in the csv files named Chicago_Crimes_2008_to_2011 and Chicago_Crimes_2012_to_2017.

4. EXPERIMENTAL EVALUATION

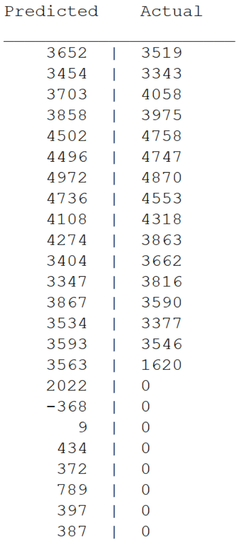
4.1 Experimental Setup

This section should include:

1. We have used the API from the Chicago Police Department's CLEAR(Citizen Law Enforcement Analysis and Reporting) system.
2. We have used 2008 to 2017 data as training set and after that we have used it for test set.
3. We have used root-mean-square-error and slope coefficients..

4.2 Experimental Results

A table of our results follows



```
Predicted    Actual
_______________________
    3652  |  3519
    3454  |  3343
    3703  |  4058
    3858  |  3975
    4502  |  4758
    4496  |  4747
    4972  |  4870
    4736  |  4553
    4108  |  4318
    4274  |  3863
    3404  |  3662
    3347  |  3816
    3867  |  3590
    3534  |  3377
    3593  |  3546
    3563  |  1620
    2022  |  0
    -368  |  0
       9  |  0
     434  |  0
     372  |  0
     789  |  0
     397  |  0
     387  |  0
```

The zeros represent months that have not occured yet. Our model did a pretty good job of predicting the months that have data. Our root mean squared error was 654 and our R-square was .8844. Clearly the 0's messed up the calculations of the model and could be removed. But as for the months with actual data, our model wasn't far off.

The slope coefficients were [0.04047659 -0.33797357 0.5207421 -0.02364527 -0.17952924 0.19964296

-0.12079136 0.01410421 0.02981009 -0.427325 0.33454637 0.91736617] and the intercept was 62.1946.

The project did an ok job. It did well predicting the months that already had data but for the months that have not happened yet, it did poor. More data could be helpful with this but the data before 2008 wasn't very reliable. We could have used the mean of previous months as the values for future months and that would've helped with better predictions, but it is the future so we don't know. Other factors may play a role such as the coronavirus which is lowering the crime rate.

5. CONCLUSIONS

We found that our model predicted the crimes per month adequately. We used 0 in the future months which threw off our calculations. Other than that, the model was good.

6. REFERENCES

[1] Data.cityofchicago.org. 2020. [online] Available at: <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2> [Accessed 11 April 2020].