Chicago Crime Rates Prediction

Anton Kamenov, Cristian Guba, Stefanos Agelastos

June 1, 2022

Abstract

By using a data-driven approach this research focuses on exploring the crime rates in Chicago from year 2012 to 2016 in order to find a matching pattern for crime prediction in the future. The primary goal of this reasearch is to predict the amount of crime, based on the date and place. Using strong machine learning models it was possible to predict the crime rate. However, due to the low accuracy of the predictions it was concluded that it is not possible to rely solely on the models produced by this research

Contents

List of Figures

1 Introduction

Crime has always been a major problem in small and big cities around the world. Throughout the years, society has been trying to reduce the number of unlawful acts committed by people. By establishing new rules and punishments for those who do not obey the law, people have found a way to reduce the crime rate. Nations have evolved their legal systems during thousands of years of trial and error, but that still does not stop citizens from commiting a crime.

There are many reasons why people still tend to participate in unlawful activities. Whether it be because of poor living conditions, psychological disorders or other, the world is always going to be a witness of crimes, no matter how scary the consecuenses of committing them are. For this reason the safety of each city is handed over to the police departments.

Police departments are another tool for lowering the amount of crimes. They operate by answering to people's calls for help and by patrolling around the city, split into designated areas. Sometimes, however, there are not enough patroling cars in the area to react quickly to the amount of crimes happening at the same time.

2 Research Question

By collecting data from one of the U.S's largest cities - Chicago, where the murder rate has remained persistently high troughout the years[3], this reasearch is going to use a data-driven approach to examine if:

- Crime is influenced by the date and time;
- Crime is influenced by the districts.

Using the analysis the reasearch is going to try to answer the question:

- Can crime rate be predicted based on the date, time and disctrict?

3 Methods

This research objectives necessitate a multi-methodological approach that integrates theory building, systems development, observation and experimentation, as described by Nunamaker et al[6]. Rather than a linear research method, this approach can be considered an agile research model, due to the continuous going back and forth between theory building, systems development, observation, and experimentaion.

3.1 Theory building

Starting with the dataset[1], the research is going to begin with data exploration, in order to see how the data looks like. Viewing the different features the dataset has and the missing values, it will help to formulate a theory that will need to be proven. After removing or extrapolating the missing values, different parts of the data can be plotted in order to gain more insight into how the data is looking.

3.2 System development

Next part is the System development, which involes creating new features to the dataset (feature engineering), converting one type of data to another (for example date and time to UNIX timestamp). After that models are selected for experimantation with the newly build data.

3.3 Observation

The Observation happens after each model is trained. By examining the result a decision needs to be made to choose which model performs the best in the given conditions with the given parameters.

3.4 Experimentation

After observation the research is continued by making small adjustments to the model parameters and the dataset in order to see what changes will take place and how will they influence the results.

4 Analysis

4.1 Data Cleanup and data exploration

With the process described in section 3.1 the research started by analyzing the content of the

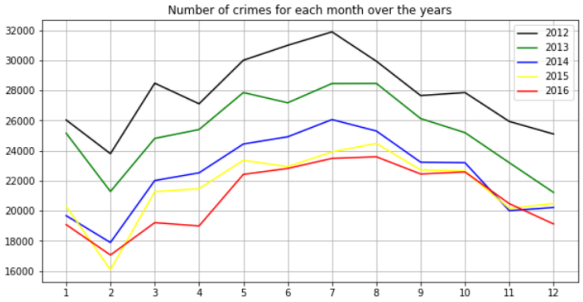

Figure 1: Number of Crimes throughout the years

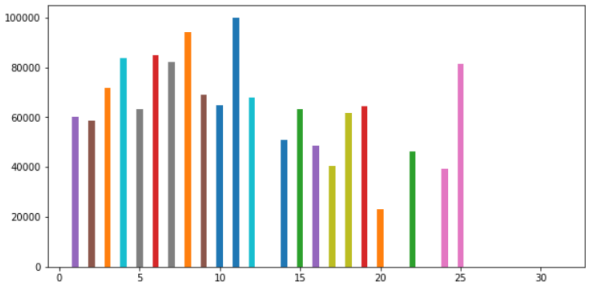

Figure 2: Number of crimes for each district

dataset. Non-relevant rows were dropped in order to gain more focus on data related to place and time. There are some specific days during which the crime rate is significantly higher, which mostly corresponds to most popular national holidays. During daytime, the crime rate in Chicago is lower and the number of the most dangerous crime types tend to increase during the night. As can be seen in Figure 1, there is a specific pattern for each year - seasonality, for how the number of crimes changes every month. There is also a trend observed, where the number of crimes is decreasing every year.

The city is divided into 22 police districts, this parameter was chosen as a feature for the location of the crime. After the graphical representation of the number of crimes in each district in Figure 2 it is possible to identify that there are districts with a higher number of crime count. This doesn't mean that a specific district should have a higher level of alarm as each district is of different size and different number of population.

For the date and time columns some aditional cleanup steps were implemented. The date and time was converted to a UNIX timestamp. This was chosen so that there is continuity in the feature, turning date and time into a discrete numerical parameter. Each timestamp represents an hour of the day for the specific day. By aggregating the number of crimes for each hour in a specific district, a new feature is created holding the count of crimes(Section 3.2).

4.2 Time Series Analysis

For the timeseries analysis and forecasting, the chosen ARIMA model is expecting a series of monodimentional, periodic data. In order to accomodate the model we removed the district feature and aggregated the crime count based on the timestamp. One can see the daily crime count for the peak months of Spring and Summer 2015 in Figure 8 of the Appendix A.2. It is noticeable that there are outliers on the first day of every month. It is uncertain if that is a systematic error, or an actual occurence so this poses an interesting problem. By choosing a weekly forecasting period, the outlier values of the first day were spread over the first week of each month (Figure 9 of Appendix A.3) smoothing out the outlier values.

4.3 Splitting the set in training, validation and testing sets

In order to train the models, tune their hyperparameters and test the results on unseen data the dataset is cleaned and split in 3 sub sets. The data is split in chronological order, where the training set is chronologically first, the validation set comes next and the test data is the last. The reason is that the research questions are trying to get answers about unknown future values – extrapolation. The aim is to train and evaluate the models in a chronological order, by learning on previous data and forecasting future uknown data. For the timeseries ARIMA model, the data is split in 75% train/validation and 25% test data, then the train/validation is further split in 75% train and 25% validation. For the rest of the machine learning models, the split is 60% training, 20% validation and 20% testing.

Timeseries Dataset:	
Size of Training set	: (147)
Size of Validation set	: (49)
Size of Test set	: (65)
Machine Learning Dataset :	
Size of Training set X	: (24147, 2)
Size of Training set Y	: (24147,)
Size of Validation set X	: (8049, 2)
Size of Validation set Y	: (8049,)
Size of Testing set X	: (8049, 2)
Size of Testing set Y	: (8049,)

4.4 Ridge Regressor model

Using the already split training set and validation set, a Ridge Regressor model was trained in order to try to predict the actual amount of crime for a speicific date in a specific period. Due to the shape of the data this experiment was not very successful. In general trying to predict the exact amount of crimes is not really useful and is prone to error.

4.5 Implementation and Parameter tuning for ARIMA

Following the methods in Section 3.2 the choice to include a classical forecasting model for timeseries seems natural, as these models are known in many cases to perform better, needing less data, as observed by Parmezanet al. [8] The Autoregresssive Integrated Moving Average (ARIMA) model is a good fit for this purpose due to the observed trend in the dataset, Figure 2. The ARIMA model is a common choice for this type of datasets, as seen in similar works by Al Balamesh et al and Eymen et al. [7] [2] The model hyperparameters are initialising the three different aspects of ARIMA, **p** for autoregression, **d** integration and **q** for the moving average. The Grid Search method was used, as described by LaValle et al, [5] and walk-forward train/validation steps were executed for each combination of the selected hyperparameter values(**??**). The Mean Square Error was used in order to evaluate and compare the predictions for each hyperparameter set (Section 3.3).

4.6 Implementation and Parameter Tuning for Decision Tree Classifier

Further in the analysis it was found that the existing data is very clustered, as can be seen in Figure 3

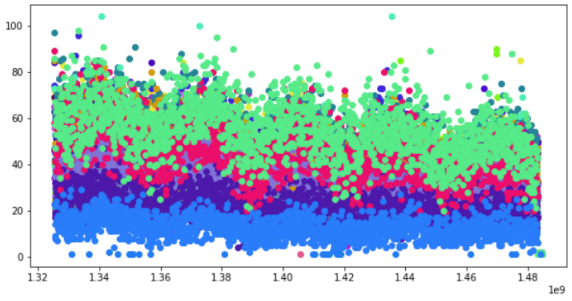

Figure 3: All crimes for each Disctict during the years 2012-2016

This type of data would indicate that using a regression model to predict the amount of crimes would not be a good fit. For this reason a classifier model would be a better fit for this dataset. In order to create and use a Decision Tree Classifier a new feature needed to be created for the dataset which would classify the amount of crimes for each day into 3 groups:

1. **Low Crime Rate - 0**
2. **Medium Crime Rate - 1**
3. **High Crime Rate - 2**

By observing the dataset it was found that the median value for the amount of crimes is 35 which will be the border between the **Low Crime Rate** and **Medium Crime Rate**. The next partition will be at 50. Every crime rate that is above 35 and below 50 will be classified as **Medium Crime Rate**. Everything above 50 and below 104(the maximum amount of crimes recorded) is classified as **High Crime Rate**.

Using the methods in Section 4.3 the new train, validation and test sets were created in order to train a Decision Tree Classifier. Grid search cross validation over specified parameter values was implemented in order to tune the decision tree model to the best performance. The hyper-parameters used for the search were:

1. **Maximum Depth - 2, 3, 5, 10, 20**
2. **Minimun Samples per leaf - 5, 10, 20, 50, 100**
3. **Criterion - gini, entropy**

This way the model will be tested with each of the parameters provided and increase the accuracy of the predictions.

5 Findings

The regression model did not perform well in this research. The overall accuracy of the model was 18% which is not really useful in any scenario.

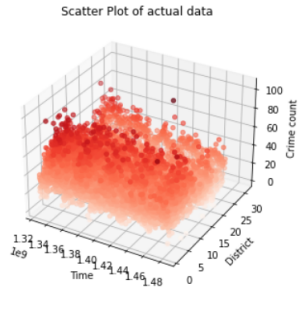

Figure 4: The data uset for testing

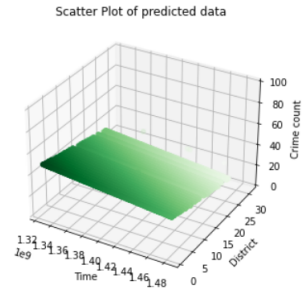

Figure 5: Predicted data

The Figures above show the differences between the actual data and the predicted data. It can be clearly seen that the model is underfitting and not generating the desired result.

The results of the Grid Search method for selecting ARIMA hyperparemeter values resulted to an ARIMA(4, 0, 0) with a Mean Square Error of 136 179 for the train/validation sets and 84 182 for the test set. This means that according to this research a simple Auto Regressive model with a lag order of 4 is enough, as values of 0 for the Integration and Moving Average are disabling these last two models. In Figure 7 of Appendix A.1 the results of the tuned model can be seen on the test data, and the actual observations.

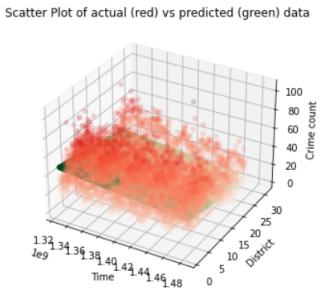

Figure 6: Predicted data vs Actual data

The result from the decission tree classifier before the tuning had an accuracy of 64.02%. The following confusion matrix shows that the model guessed correctly mostly the crime rate classified as **Low**.

Predicted Rate	0	1	2
Actual Rate			
0	4610	651	45
1	1695	406	191
2	273	41	137

Further analysis in the classification report shows that the precision is not so high for the Medium and High Alarm which are more important in therms of accuracy.

	precision	recall	f1-score	support
0	0.70	0.87	0.78	5306
1	0.37	0.18	0.24	2292
2	0.37	0.30	0.33	451
accuracy			0.64	8049
macro avg	0.48	0.45	0.45	8049
weighted avg	0.59	0.60	0.60	8049

By tuning the Classifier using Grid Search it turned out that the following parameters are the best for this type of classification.

- Maximun depth = 10;
- Minimun samples per leaf = 100;
- Random state = 42;

The accuracy went up to 0.71%.

The classification report shows that although the accuracy went slightly up, the most important predictions(Medium and High Crime rate) are still low on the scale.

	precision	recall	f1-score	support
0	0.79	0.87	0.83	5306
1	0.53	0.42	0.47	2292
2	0.37	0.30	0.33	451
accuracy			0.71	8049
macro avg	0.56	0.53	0.54	8049
weighted avg	0.69	0.71	0.70	8049

6 Conclusion

Based on this research, it can be concluded that time and place are factors which influence the number and type of crimes in a specific district, which could help police to distribute their resources better by planning where and how many officers should be deployed to patrol. Is this information enough for an artificial intelligence model to predict where and which type of crime could happen? Clearly not, as the behavior of human beings is not so easy to predict, usually it is influenced by a lot of external factors which are very important and not included in our data. Covid-19 was a factor that changed people's behavior and their mental health [4]. Therefore, crime amount and their types changed significantly during lockdowns. Restrictions also caused limitations in people's movement. Using such models to predict crime poses some serious ethical questions, as this is a very sensitive subject. Current generations tend to prioritize human life and the job of the state is to protect it.

Bibliography

[1] Spiros Politis (2018). "Chicago crimes 2001-2018 (November)". In: (2001-2018).

[2] Abdurrahman Eymen and Ümran Köylü. "Seasonal trend analysis and ARIMA modeling of relative humidity and wind speed time series around Yamula Dam". In: *Meteorology and Atmospheric Physics* 131.3 (2019), pp. 601–612.

[3] Matthew Friedman, Ames C Grawert, and James Cullen. *Crime Trends, 1990-2016*. Brennan Center for Justice at New York University School of Law New York, NY, 2017.

[4] Dae-Young Kim and William P McCarty. "Exploring violent crimes in Chicago during the COVID-19 pandemic: do location, crime type, and social distancing type matter?" In: *Journal of Crime and Justice* (2021), pp. 1–16.

[5] Steven M LaValle, Michael S Branicky, and Stephen R Lindemann. "On the relationship between classical grid search and probabilistic roadmaps". In: *The International Journal of Robotics Research* 23.7-8 (2004), pp. 673–692.

[6] Jay F Nunamaker Jr, Minder Chen, and Titus DM Purdin. "Systems development in information systems research". In: *Journal of management information systems* 7.3 (1990), pp. 89–106.

[7] Oluwakemi Ololade Odukoya et al. "Epidemiological trends of coronavirus disease 2019 in Nigeria: from 1 to 10,000". In: *Nigerian Postgraduate Medical Journal* 27.4 (2020), p. 271.

[8] Antonio Parmezan, Vinícius Alves de Souza, and Gustavo Batista. "Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model". In: *Information Sciences* (Jan. 2019). DOI: 10.1016/j.ins.2019.01.076.

A Appendix

A.1 Figure 7 - ARIMA test set results

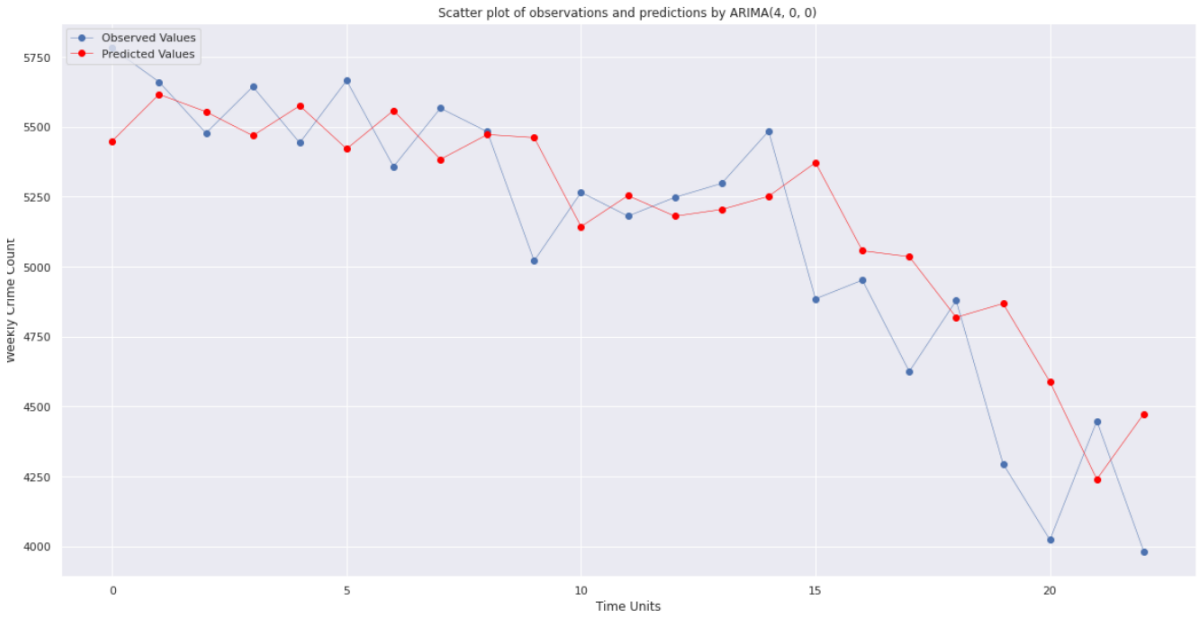

Figure 7: Test set observations and ARIMA forecasting

A.2 Figure 8 - Spring and Summer 2015 daily Crime count

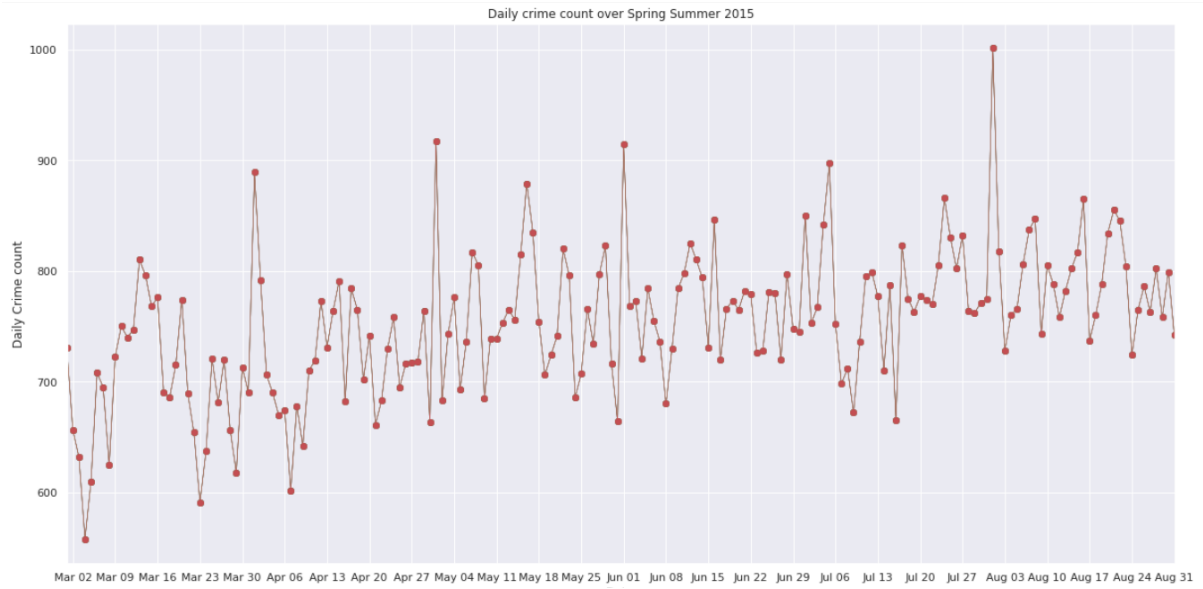

Figure 8: Spring and Summer 2015 daily Crime count

A.3 Figure 9 - Spring and Summer 2015 weekly Crime count

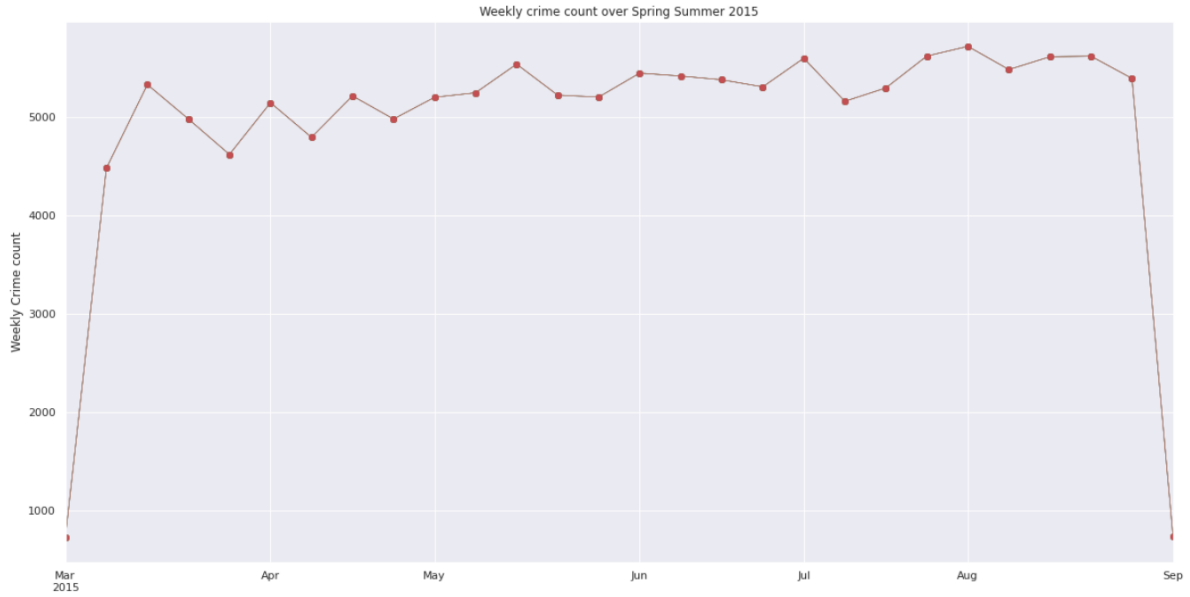

Figure 9: pring and Summer 2015 weekly Crime count