Institute of Technology Carlow

Data Science

Crime Analytics & Prediction in Chicago

Project Specification

Ger Dobbs C00196843 Karl Redmond C00196815 David Kelly C00193216

Lecturer: Greg Doyle

Submission Date: 17/10/2017

Table of Contents

Introduction

This 4th year Data Science project has been completed with a view to analysing and predicting crime in the chicago area. It will begin by explaining the purpose and goals of the project. This will include an explanation as to what predictive analysis is and how it relates to the goals we wish to achieve, which will be explained in the learning outcomes. On commencing this project, it was important that we set out our strategy, in order to reach our goals and learning outcomes. We will then define and map our scope in order to ensure the project is completed successfully and on time. This will outline both the responsibilities of each project owner and what is to be included in the project. The tools we researched and ultimately chose to use are explained. An outline will be given of the reasons for the decisions we made.

Purpose

"Predictive policing is the use of analytical techniques to identify promising targets for police intervention with the goal of preventing crime, solving past crimes, and identifying potential offenders and victims" [1]

Predictive policing can be used to forecast the location and time of a crime occurring. It can also be used to predict the perpetrator and the victim. The approach taken in this project is to try to predict where and when a crime will take place. The ultimate goal is to use data analysis, to examine if historic crime data can be used to predict crime patterns and pre-empt where and when particular crimes will occur. This could lead to the ability to identify areas of high risk and ultimately lead to decisions being made on policing policies. It is not intended as a replacement for other policing strategies but as a complement to existing techniques. It would also be more useful as a crime detection tool as crime occurs in real time.

The results conveyed from the analysis of the chosen dataset can be used in crime prevention in two ways. The first is in real time, where a crime has been reported and we have identified that there is a very strong risk of another crime being committed in the same area in a certain time period. The second crime prevention technique is where the data analysis of historic crime patterns predict that certain locations are susceptible to particular crimes during a specific time period. For instance, a burglary has been reported in a specific area. There is now a risk that more burglaries could take place within a defined area and within a certain time period. If another crime then occurs in the same area, and within a certain time period, then this will

further increase the chances of more burglaries in the area. This could then lead to extra police patrols in the affected area.

"That doesn't mean police can prevent every kind of crime—only the more predictable kinds, like burglary or auto theft," said Beam. "And even then, some robberies could be truly random. But predictive policing could reduce crime on the margins, according to the UCLA researchers."[3]

Along with the ability to predict crime taking place in in real time, the data set could also be used to point out crime hotspots. These hotspots could be linked to other factors such as the time of year or events occurring in the area. The results of such analysis could be useful in an overall policing strategy, depending on the results found. Analysis of results obtained in this instance, would be more appropriate for crime prevention, as opposed to crime detection.

The goal of this project is to analysis the data set for patterns. It is hoped the results will provide us with important information which could aid the prevention and detection of crime. This could then be used, along with other policing strategies to direct policy. It would not be intended to replace the use of police intuition, experience or current practices.

Using data to assist in the prediction of crime can have its drawbacks. It is heavily reliant on the quality of the data used. Using a dataset that has a degree of accuracy may lead to results which are skewed and therefore to a policy that is not truly reflective of the needs. For example, data on unreported crimes is not recorded in our dataset and may have an adverse effect on the results and therefore the policing decisions undertaken. An assumption is also made that all arrests are recorded and that all arrests are justified.

"The unequal treatment of minorities in the criminal justice system is one of the most serious problems facing America in the new millennium."[2]

In recent times there have been accusations of a concentration of policing on certain demographics. Nationwide protests have been organised by the BlackLivesMatter movement. These protests are due to the perceived unequal treatment of black people by police authorities in the United States. The veracity of these claims will not be of interest to this project, except to say it may distort the data slightly. This distortion may have a knock-on effect in analysing this data for the purpose of predictive policing. If policing is concentrated in the wrong areas, it may lead to more protests and increase the perception of police bias. In this regard the data may be slightly biased and distorted but with such a big dataset this bias will be reduced.

Goals

The project will concentrate on a dataset extracted from Kaggle. It has over 6 million entries for the period 2001 to present. We chose this dataset because of the volume of data recorded. Extending this dataset to an extended period of time would further help alleviate any possible bias. In recent times, due to very public instances as outlined above, there is a heightened awareness of possible police bias. The result should be an improvement in data veracity, if any inaccuracies do exist, thus improving the accuracy of results provided by the algorithms used in this project.

"The prediction of future crime trends involves tracking crime rate changes from one year to the next and used data mining to project those changes into the future." [4]

It is proposed that the continued analysis of crime data will lead to patterns and predictions will uncover previously unknown trends in the commission of crime. Although the data set we have used is large, it is for a relatively short period of time, ie. three years. Extending the period analysed, would in the future, provide more detailed results. Predicting where and when crime will occur is not a new phenomenon. Police forces have been doing it for years. However, what is new, are the tools and methods available to carry out such predictions. The amount of data available has grown exponentially. As the data sets on crime grow and as their veracity improves, then the analysis of such data will make any results gathered more accurate. This project aims to use the previously mentioned data to discover any patterns in the crime figures. We will then interpret how useful the analysis can be in predicting the committal of crime.

"So, does it work? According to the BBC, in instances where specific patterns are exhibited, it's helping to cut down on crime. In fact, Manchester police noticed a 26.6% decline in burglaries in 2011." [3]

Learning Outcomes

This project will present us with the opportunity to gain experience in a field that we have little prior knowledge of. Through the development of this project we hope to gain a fundamental knowledge base and skillset of data science, analysis of large datasets and associated technologies.

We will have the opportunity to learn and use new technologies, such as Anaconda and Jupyter, for collaboration and displaying our data. We will further develop our skills in Python and gain insight and a working knowledge of the extensive range of Python data mining libraries.

Lastly we hope this project will enable us to apply our computer science and programming knowledge to perform analysis, hopefully make predictions and maybe prescribe countermeasures to a fundamental problem in society.

Project Strategy

In order to successfully reach our projects goals and obtain our anticipated learning outcomes we defined a strategy to guide us through our data mining process and ultimately towards knowledge discovery. Following our initial research, we will adhere to the Knowledge Discovery in Databases process (KDD). The following describes our applications of different techniques towards crime data knowledge discovery.

Research

Initially, research will provide us with a foundational knowledge to confidently select a domain and data set. We began by searching for similar studies or projects that had been performed in the domain of crime data. We needed to understand the scope and outcomes of previous projects in order to correctly gauge our deliverables, goals and anticipated learning outcomes.

The most significant application of data mining in crime analysis and prediction can be seen in Los Angeles, California. From 2014, the Los Angeles and Santa Cruz Police Departments have been using big data algorithms to predict where crime is likely to occur. According to Mark van Rijmenam [5], the success of their efforts is confirmed with a 33% reduction in burglaries, 21% reduction in violent crimes and a 12% reduction in property crimes.

PredPol Inc. is the company responsible for the crime prediction software. The company used an algorithm used to predict earthquake aftershocks, modified it then began feeding it crime data. As a result, PredPol is cited as:

> "…a leading crime data mining & predictive policing solution..." [6]

According to PredPol, the software can predict where crimes are likely to occur down to 500 square feet. And as a result, command staff and crime analysts using the software are 100% more

effective in their roles when compared to those using traditional crime hotspot mapping, therefore law enforcement have twice as many opportunities to deter and reduce crime.

Concluding our initial research, we can develop an understanding of the application domain, the relevant prior studies, projects and knowledge; and, the anticipation of the end-user.

Selection

We chose a dataset of crimes in Chicago. The dataset was made available by the City of Chicago and distributed on Kaggle Datasets website. The full dataset contains 6,000,000 records (rows) and is available in comma-separated value (csv) format.

This dataset reflects reported incidents of crime that occurred in the City of Chicago from 2001 to present and is current to the previous week. It is extracted from the Chicago Police Department's CLEAR system (Citizen Law Enforcement Analysis and Reporting) [7]. In order to protect the victims' privacy, address information is displayed at a block level and precise crime location is shifted from the actual location for partial redaction but remains within the same block.

Each dataset record contains 22 attributes (columns), including: ID, Case Number, Date/Time, 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.

First, we will focus on crimes committed in 2014 and 2015, with the aim of generating crime predictions that correlate to the 2016 records. From there we may expand our sample to additional years. We anticipate that our preliminary variable subset will include: Date, Time, IUCR, Primary Type, Location, Arrest, Location

Pre-processing

We anticipate a certain degree of cleaning will be required before processing the data, such as removing outliers. Although we are dealing with a substantial dataset, strategies for dealing with missing data fields will need to be explored. According to Rachel Margolis, Phd [8] if records with missing data are deleted we risk reducing the sample size and lowering statistical power with biased estimates. Also, if we impute missing data, there is risk of biased estimates and inadequate imputations.

Data Mining

We will use cluster analysis for our initial investigation. We aim to partition the crime data into groups based on similarities such as time, location, date or the type of crime committed. Our hope is that this will allow us single out or distinguish varying crime hotspots in the City of Chicago and identify characteristics of these hotspots. Also, we may be able to identify outliers within the dataset.

Scope

The scope of this project relevant to the data set used is from the period 2014 - 2017. This data was extracted from, and can be viewed at :

https://www.kaggle.com/amunnelly/crime-in-chicago/data

We have limited the scope of the data used to this period due to the volume of the data. It is planned to expand this scope further should the thorough analysis of this data return interesting results. The analysis will be carried out as per details as outlined in our strategy.

As project owners, we plan to divide and share the workload evenly to ensure all deliverables are returned on time.

Deliverables

Data Mining is an analytic process designed to explore data in search of consistent patterns between variables, and then to confirm the findings by applying the detected patterns to new subsets of similar data.

There are a number of deliverables or goals associated with data science, including [9]:

- Prediction (predict a value based on inputs)
- Classification (e.g., spam or not spam)
- Recommendations (e.g., Amazon and Netflix recommendations)
- Pattern detection and grouping (e.g., classification without known classes)
- Anomaly detection (e.g., fraud detection)
- Recognition (image, text, audio, video, facial)

- Actionable insights (via dashboards, reports, visualizations)
- Automated processes and decision-making (e.g., credit card approval)
- Scoring and ranking (e.g., FICO score)
- Segmentation (e.g., demographic-based marketing)
- Optimization (e.g., risk management)
- Forecasts (e.g., sales and revenue)

The goal of this project is to make predictions on the possible occurrence of crime in a certain location and at certain times. For this reason, the deliverable of this project will be that of prediction.

Predictive data mining is the most common type of data mining, and consists of three stages:

- Initial exploration
- Model building or pattern identification with validation/verification
- Deployment (i.e. the application of the model to new data in order to generate predictions).

Taking these stages, our initial exploration will be based around a dataset titled “Crimes in Chicago” which was retrieved from kaggle.com. This data consists of various crimes committed from 2012-2017.

We will build our model around 2014-2015, with validation/verification coming from 2016.
The model will then be applied/deployed to 2017 to generate predictions.

Functional Interfaces

There are a multitude of tools which can be used to explore or apply “data science” algorithms to data sets or “big data”. The following graph outlines the most popular tools as of 2017.

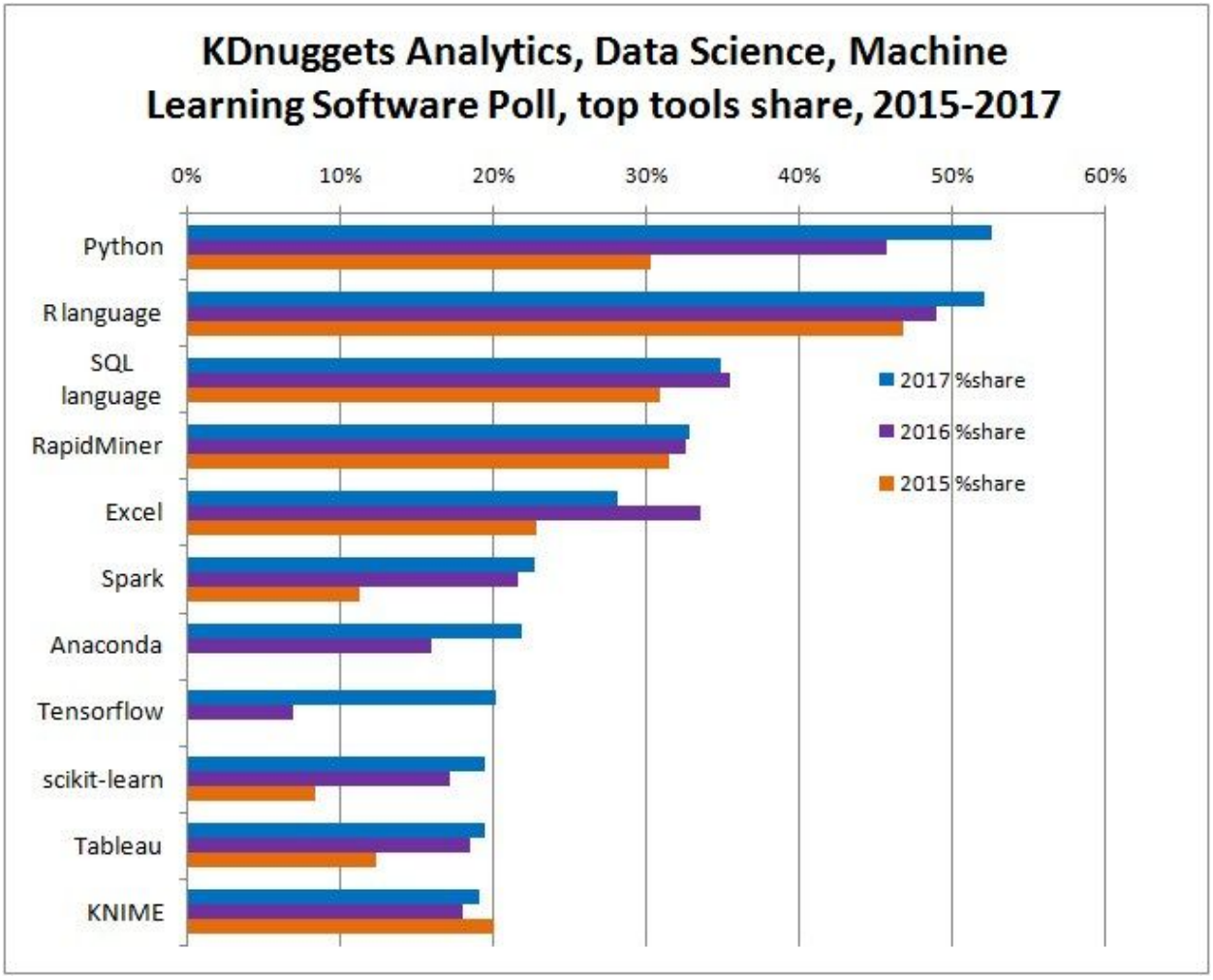

Fig. 1. [10]

As our team has previous experience with Python, together with the fact that Python usage for data analytics is on the rise, we decided to investigate other tools which could be used in conjunction with Python to create a useful interface for data exploration. Outlined below are some useful libraries and a popular IDE called Jupyter.

Python

"Python is powerful... and fast; plays well with others; runs everywhere; is friendly & easy to learn; is Open." [11]

The above statement is taken from Pythons homepage, and rings true when putting it to use. It is a versatile language which can be used for many different tasks including, but definitely not limited to, data science.

Python has quickly become an indispensable tool used by data scientists (see Fig. 1.), which topped the list of data analytic tools used in 2017, as opposed to 2015 where it came in fourth, and is up 15% in usage since 2016 (see Fig. 2.).

Table 1: Top Analytics/Data Science Tools in 2017 KDnuggets Poll

Tool	**2017 % Usage**	**% change 2017 vs 2016**	**% alone**
Python	52.6%	15%	0.2%
R language	52.1%	6.4%	3.3%
SQL language	34.9%	-1.8%	0%
RapidMiner	32.8%	0.7%	13.6%
Excel	28.1%	-16%	0.1%
Spark	22.7%	5.3%	0.2%
Anaconda	21.8%	37%	0.8%
Tensorflow	20.2%	195%	0%
scikit-learn	19.5%	13%	0%
Tableau	19.4%	5.0%	0.4%
KNIME	19.1%	6.3%	2.4%

Fig. 2. [10]

Python Packages

There are thousands of packages available for Python, some of the most useful are described briefly in the following:

- NumPy is essential in scientific computing using Python. It utilizes a number of useful functions, which can be used for fast operations on arrays.
- SciPy is built upon NumPy and contains modules for optimization, linear algebra and interpolation.
- Pandas is used to work with tabular data, for quick manipulation and visualization of the data. Some examples of use are deleting and adding columns, handling missing data and grouping by functionality.
- Matplotlib is a 2D plotting library, which can be used in the jupyter notebook and web application servers. It can be used to generate plots, histograms, power spectra, bar charts and scatterplots.
- Seaborn focuses on the visualization of statistical models, and provides an interface for drawing graphs.
- SciKit-Learn provides a number of learning algorithms for machine learning facilitation. The package is built on top of SciPy and makes heavy use of its maths operations.

Jupyter

"One of the most significant advances in the scientific computing arena is underway with the explosion of interest in Jupyter(formerly IPython) Notebook technology". [12]

Jupyter Notebook is a web application based on server-client structure. Starting the notebook server is a simple process, simply typing jupyter notebook on the command line will start the server, and allow the user to use the notebook on localhost. The Jupyter interface allows users to run 'code blocks', and uses a form of programming called 'literate programming'. Literate programming is a software development style which utilizes the approach of writing your thoughts in English, supplemented by mathematical equations, and subsequently include code blocks. These blocks are individually runnable, allowing on the go manipulation and tweaking, with the results instantly viewable.

Jupyter Notebook is easy to use and provides you with an interactive environment, supporting HTML, video, code and data visualization libraries. This ultimately allows you to show your graphs in the same document as your runnable code.
These documents can be exported in PDF or HTML format, which lends itself to the presentation of the work carried out by your research and algorithms.

Anaconda

Most of the aforementioned libraries, as well as Python and Jupyter, are included in the Anaconda distribution.
The distribution aims to provide everything you need for data science. Using this platform allows teams to quickly get up and running, cutting back on environment set-up time. It can be run on most popular operating systems, including Windows and Mac OS.

Conclusion

Our goal is to investigate whether we can make accurate predictions as to where and when a crime could occur in the city of Chicago, Illinois. We will do this by utilizing the functionality of Jupyter, in conjunction with Python code, to run a predictive algorithm on the data set titled "Crimes in Chicago". By taking the data from years 2014-2015 to come up with our model, and verifying the model on year 2016, we will hope to produce predictions for year 2017.

References

[1] Rand Corporation. (2017). Predictive Policing, [online], available: https://www.rand.org/pubs/research_reports/RR233.html [accessed 12 October, 2017].

[2] Clarence M. Dunnaville, Jr. (2000). Unequal Justice Under the Law — Racial Inequities in the Justice System. [online], available: http://www.vsb.org/docs/valawyermagazine/dec00dunnaville.pdf [accessed 12 October, 2017].

[3] GRAYMATTER. (2015). Using Big Data to Predict Crime Patterns, [online], available: http://graymattersystems.com/big-data-crime-patterns/ [accessed 12 October, 2017].

[4] The International Journal of Engineering And Science (IJES). (2012). http://www.theijes.com/papers/v1-i2/AJ01202430247.pdf [accessed 12 October, 2017].

[5] Mark van Rijmenam. (2017). The Los Angeles Police Department Is Predicting and Fighting Crime With Big Data, [online], available: https://datafloq.com/read/los-angeles-police-department-predicts-fights-crim/279 [accessed 12 October, 2017].

[6] PredPol. (2017). *Policing in the 'Big-Data' Crime Prevention Era*, [online], available: http://www.predpol.com/data-mining-crime-predictions/ [accessed 12 October, 2017].

[7] Kaggle. (2017). Crimes in Chicago, [online], available: https://www.kaggle.com/currie32/crimes-in-chicago [accessed 12 October 2017].

[8] Rachel Margolis. (2013). *Dealing with missing data,* [online], available: http://rdc.uwo.ca/events/docs/presentation_slides/2012-13/Margolis-MissingData2013.pdf [accessed 12 October 2017].

[9] KDNuggets (2017). What is Data Science, and What Does a Data Scientist do? [online], available: https://www.kdnuggets.com/2017/03/data-science-data-scientist-do.html [accessed 17 Oct 2017].

[10] KDNuggets (2017). New Leader, Trends, and Surprises in Analytics, Data Science, Machine Learning Software Poll [online], available: http://www.kdnuggets.com/2017/05/poll-analytics-data-science-machine-learning-software-leaders.html [accessed 12 October, 2017].

[11] Python Software Foundation (2017). Python.org About Page[online], available: https://www.python.org/about/ [accessed 12 October, 2017].

[12] Unidata Online Python Training. Why Python and Jupyter Notebooks?[online], available: https://unidata.github.io/online-python-training/introduction.html [accessed 13 October, 2017].

Research Documents

http://www.ipcsit.com/vol6/26-E049.pdf

http://www.theijes.com/papers/v1-i2/AJ01202430247.pdf

http://www.predpol.com/data-mining-crime-predictions/

https://datafloq.com/read/los-angeles-police-department-predicts-fights-crim/279

https://www.theguardian.com/cities/2014/jun/25/predicting-crime-lapd-los-angeles-police-data-analysis-algorithm-minority-report