



COMP8410-Data Mining

Semester 1, 2022

The Ethics and Social Impact of Applying Data Mining in Predictive Policing

Course Code: COMP8410

Course Name: Data Mining

Lecturer's Name: Kerry Taylor & Pouya Omran

Tutor's Name: Cheng Xue and Zhangcheng Qiang

Date/time due: Monday Week 4 (14th of March 2022) / 9am

Student number	Student name	Contribution	Signature
u7167784	Man Jin	100%	<i>Man Jin</i>

Table of Contents

1. Project Description	3
1.1 Aims	3
1.2 Methods	3
1.3 Data source and nature	3
1.4 Authority	4
1.5 Expected impact	4
1.6 Other uses of the results	5
1.7 Ethical considerations	5
2. Ethical Aspects	6
2.1 Potential ethical issues	6
2.2 Biased decisions	6
2.3 Individual privacy	7
2.4 Public interest	7
2.5 Analysis of the issues	8
3. Recommendations	8
3.1 Technological solutions	8
3.2 Procedural approaches	9
3.3 Governance approaches	9
3.4 Educational approaches	9
3.5 Other ideas	10
References	11
Appendix	12

1. Project Description

1.1 Aims

In many situations, the focus of resource allocation is to do as much work as possible with current resources or even fewer resources (McCue, 2006). The goal of the police department is the same. When the number of crimes rises, we subconsciously believe that we need to increase the number of police and deploy them in large numbers. However, with the decreasing police human resources and law enforcement budget, simply increasing the police force and accelerating the response speed to cases have faced a significant bottleneck in improving the effect of maintaining social security and combating and preventing crime. The aim of this project is to predict the "crime category" of a region of the Republic of Poland and the correlation between trends, seasonality, and holidays with crime frequency (Borowik, Wawrzyniak, & Cichosz, 2018).

1.2 Methods

The police resources are allocated to specific crime areas efficiently by using a decomposable time series model and successfully implemented in Facebook Prophet (Prophet prediction model). The progress of data mining is data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, knowledge presentation (Lecture Notes, 2022).

In this project, Borowik et al. (2018) share in their study that crime records are recorded according to time, which is time series data. Therefore, classification algorithm and time series algorithm can be used to model crime records. The steps of modeling process include data preprocessing, modeling, model evaluation and model application. The format of the source data is not suitable for applying the data mining algorithm immediately, thus, the data that does not conform to the format should be deleted (Fig.3), and then implement the model in a Python package called Prophet, the analysis determine the trend and seasonality of various criminal events in the region. The decomposition of the time series model includes three main parts, which are trend, seasonality, and holidays (Fig.4).

- Classification algorithm is categorizing each event or object and then predicting the future trend according to historical data. For example, all crime cases are categorized according to their location, and the categorization results in an analysis of which areas have a high incidence of cases, so that police deployment is increased as appropriate
- The time series model is finding the temporal correlation between incidents. In predictive policing, the temporal correlation of cases in terms of trends, seasonality or holidays is identified to predict the time of occurrence.

1.3 Data source and nature

The source of the data is the data collection system of a regional police department in the Republic of Poland, with approximately 1.2 million pieces of crime data. The time ranges from 1 July 2013 to 30 June 2016. The data records include the following nature information, the number of crimes reported, the district, the date, the time, the type of crime and the location of the GPS coordinates. The data is then aggregated according to the needs of the training model (Borowik et al., 2018).

1.4 Authority

Crime data records are collected by the police data collection system, which contains data from open sources, citizen reports and information collected and uploaded by the police. While the public is wary of the government's unfettered access to and use of data, many are concerned about the powerful and unsupervised data mining capabilities of police agencies and believe that even without access to back-end data, the police can still access a wealth of information about individuals' privacy.

1.5 Expected impact

The study presents the results of six categories of data that police interventions, burglary and theft, hooliganism, road traffic crimes, other criminal offences, and detention (Borowik et al., 2018).

All six types of events have seasonal characteristics. Most road traffic crimes occur in warm months and increase during Christmas and Halloween holidays (Fig.1d). Summer is also a high-frequency period for other criminal offences. In the burglary and theft category (Fig.2b), most incidents occurred around 10 a.m. and 3 p.m. because the criminals took advantage of the knowledge that the owner was not in the house or apartment. What is more, after training data in the Prophet model, then "predict" the possibility and degree of crime in separate places according to the date, the time, the type of crime, and display it on the electronic map in a visual way, which relates to the GPS system of police officers.

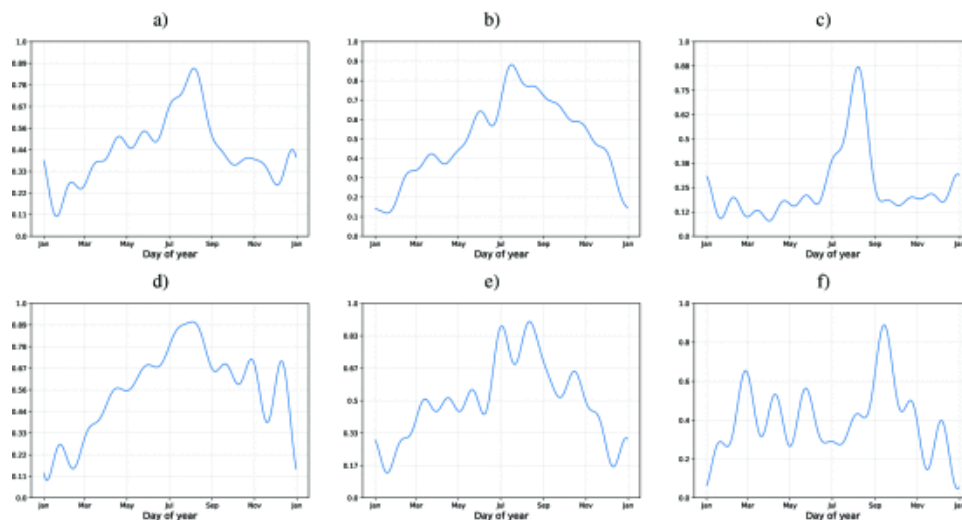


Fig.1

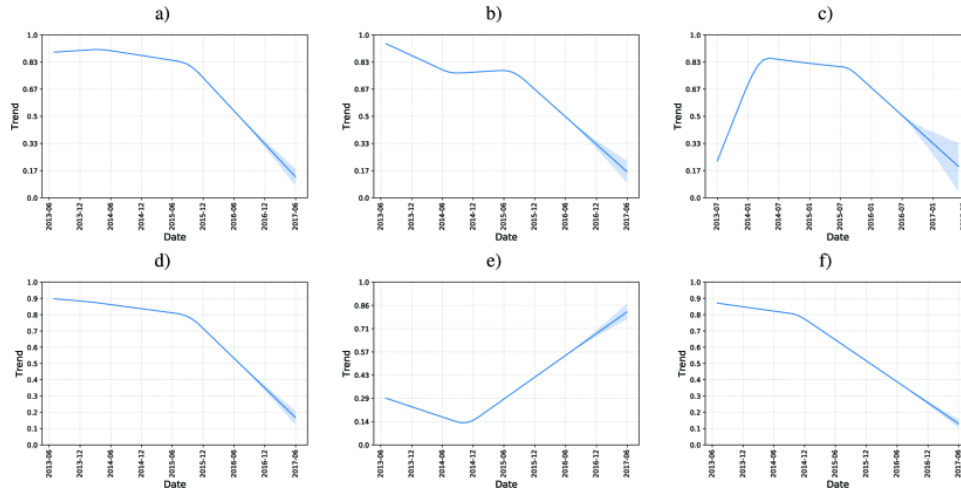


Fig.2

Therefore, through the upload and analysis of these data, the crime situation and criminal portraits of the whole region may be mastered by the police, so that the police can set the priority of criminal events, more effectively deploy the police force, and complete the task of combating crime. Additionally, since the application of the big data crime prediction model, the ratio of various local illegal and criminal acts has decreased visibly.

1.6 Other uses of the results

Although the purpose of this paper is to demonstrate the types, regional, and time-seasonality of classification and time series algorithms in predicting crime, there are uses in the law enforcement field, such as identifying crime hotspots and creating crime profiles or criminal portrait. In addition, factors associated with crime can be identified, such as the relationship between crime and unemployment rates, previous criminal records, and alcohol sales.

1.7 Ethical considerations

Spiekermann (2012) argues that it is not true that privacy and security are in opposition to each other, and that big data and smart data actively construct privacy and security. This may be true, but it is not evident in the fight against organized crime. In both the military and humanitarian spheres, effective predictive policing tools are designed to control them: to discover and access as much personal information about individuals and organizations as possible. Therefore, predictive policing automatically makes predictions about the type of crime and the characteristics of criminals. Some people could think there is a good point that the police use high-tech technology to protect cities. On the other hand, some people fear that digital power will lead, as in some science fiction movies, to a range of moral and social impacts because human based predictive policing is the most controversial type.

2. Ethical Aspects

2.1 Potential ethical issues

Although law enforcement agencies and police departments have a positive attitude towards the prospect of predictive policing, some scholars and scientists have raised concerns about the application of data mining technology to predict criminal behavior. In this project, we use data mining classification algorithm and time series algorithm to predict the types and trends of crimes in a district of the Republic of Poland, and obtain the suspect's GPS, race, and time to commit crimes, and speculate on the choice of lifestyle of suspects. This will bring a series of potential social and moral impacts.

Predictive policing algorithms usually have many data sources. The more data, the greater the noise. Big data may interfere with police prediction and affect the police to make correct decisions (Momsen & Rennert, 2022). Predictive policing does not rely solely on intuition, experience, and knowledge, but uses dynamic mathematical model and carries out prediction through computer operation. However, in fact, the police prediction using data mining technology still has the phenomenon of inaccurate prediction. This is because the object of police prediction is people, who have their own feelings and will. If machines are used to predict people's behavior under specific circumstances, human privacy will be violated and there will be no freedom. Moreover, even powerful technology cannot deeply predict people, such as criminal motivation or criminal tendency.

The criminal record data used in predictive policing may be biased (Momsen & Rennert, 2022). Heaven (2020) emphasized that the arrest rate can easily affect the prediction algorithm. Human bias has been incorporated into these tools because machine learning models are trained based on biased police data.

In the public interest, although the predictive algorithm can effectively predict the GPS of criminal events, to deploy the police force in a specific range to protect the safety of the community. However, if only one community is targeted, the relationship between the police and the community will become not close enough. Because the target orientation method used by predictive policing is not conducive to police and public relations (Hustedt, 2016), and if the police carry out surveillance or intervention measures for a specific group, it will not be conducive to citizens' trust in the police and the state, thus reducing the public interest. Next, we will discuss the moral impact of predictive policing around specific reference.

2.2 Biased decisions

Predictive policing software brings bias, not solves it. There are seven principles mentioned for Algorithmic Transparency and Accountability by Association for Computing Machinery (ACM), which includes awareness, access and redress, accountability, explanation, data provenance, auditability, validation, and testing. This document states that algorithms and analysis may be opaque, which will lead to biased or incorrect results ("Statement on Algorithmic Transparency and Accountability", 2017).

Algorithms are also used in this project. Classification algorithms and time series algorithms use past performance to predict future performance to assign polices to areas with more crimes.

Having more police in that area means finding more crimes, so the data are biased, which violates the data provenance principle. This principle points out that the builder of the algorithm should describe the way of the training data set. Nine (2022) states that the data should be unbiased from the beginning if the algorithm wants to predict based on the previous data. It is exceedingly difficult for any data set, especially in the field of law enforcement and police. Therefore, the basic data of predictive policing itself may have been polluted by biases and discriminatory practices in police law enforcement, and the data set trained on this basis is also a repetition of previous mistakes.

2.3 Individual privacy

Predictive policing may pose a threat to personal privacy. In the era of small data, the data collected by people are single purpose. When collecting data, privacy can be prevented from being leaked in the use or reuse of data through ambiguity and anonymous. In addition, it is difficult to establish a relationship between data, so it is difficult to find the secrets hidden between data. However, in the big data era, all kinds of big data are preserved forever. These data can be used repeatedly. From the perspective of data that be used for single purpose, privacy information can be shielded after being blurred or anonymous, but big data can connect small data that are not connected. Data mining technologies can associate various information fragments, which may mine again the original information. Therefore, for big data technology, the traditional methods such as fuzziness and anonymity to protect privacy are ineffective.

According to this project, the problem of individual privacy is not clear to what extent citizens' digital footprints are private and whether they can be used unconditionally. The digital footprint of citizens can be critical or justice investigations, race, and lifestyle. Violating personal privacy may occur in the data collection process without the consent of citizens or without the knowledge of citizens. Guide to data analytics and the Australian Privacy Principles (APP) (2018) states that when personal information is collected by relevant departments, APP 5 needs to notify certain matters about itself or ensure that individuals understand these matters.

Therefore, maintaining trust between citizens and the government is important because a balance should be struck between the use of big data and citizens' privacy. Without clear boundaries, citizens may have a deep distrust of the government because they do not know whether and to what extent they are monitored.

2.4 Public interest

The management of predictive policing technology may also bring unexpected consequences. Ideally, it can reduce the crime rate in the region and improve the quality of life. It can also stimulate the capital investment of the city, which is beneficial to the public interest.

However, the project uses algorithmic training data to identify individuals most likely to participate in criminal activities, as well as crime categories and predict crime trends and seasonality. Due to the opacity of the algorithm, it is not completely in an ideal state. According to ACS Code of Professional Conduct (2014), the public interest is the most important and it should be used to resolve all conflicts related to public interests include public health, safety, and environment.

Therefore, the application of predictive policing in a certain area of the city will also affect the public interests of the area and its residents. When an area continues to monitor criminal activities, the relationship between the police and the community becomes not close enough. On the other hand, another region is booming with capital investment, which will lead to cross racial unequal social strata. Thus, it endangers the normal operation of society and is not conducive to the public interest.

2.5 Analysis of the issues

The algorithms used in this project are biased, violate people's privacy and cause unequal social strata to the public interest. In fact, the problems affect each other. How to balance them is worth thinking about. The root cause is data. If the data is inaccurate, the decision-making based on data will not be accurate. Most police departments do not have appropriate processes and rules to support the informed use of data. In addition, we need a comprehensive data collection method that minimizes bias. Therefore, people need to reduce avoidable wrong judgments from the source, and do reflection regularly, rather than predicting based on inaccurate data repeatedly, resulting in a vicious circle.

3. Recommendations

3.1 Technological solutions

Some solutions can be provided from a technical point of view. The most critical point is the accuracy of the input data because the quality of prediction tools depends on the data they provide. In other words, if the data entering the model is defective and biased, the predicted results are meaningless.

The data collected by the police are opaque, incomplete, and vulnerable to prejudice. Moreover, predictive policing relies on historical crime data, which may be incomplete, such as unreported criminal records. It may also be biased, such as drug-related crimes. Therefore, how to improve the accuracy of police prediction needs to start from the root data collection.

The first technical solution is an interactive database between the police and the people open to the society. Through the technical analysis of the police's violent law enforcement, complaints and other data, the bad law enforcement data are excluded. The second technical solution is to use interpretable machine learning technology. By interpreting the model, human trust in the model can be achieved to create more secure and reliable applications (Du et al., 2019). The necessity of interpretability is that the theory behind it is human centered. According to the model training cycle, the method to realize interpretability is divided into three stages: before model training, during training and after training. Try to understand the data as much as possible before training. During the training, we need to create interpretable models, such as using decision tree and other technologies. After training, it is a step related to decision-making. Sensitivity analysis and gradient based methods are often used to interpret data.

Therefore, before any innovative technology is used by law enforcement agencies, there must be a thorough public debate and a rigorous and independent assessment of the effectiveness of the technology and its impact on the community.

3.2 Procedural approaches

The police forecasting process can also be optimized. First is data collection stage: all policing prediction technologies rely on a large amount of data, and the integration of different data resources is the basis of policing prediction. In 3.1, it is mentioned to create an interactive database between the police and the people open to the society, which will effectively clean the data before training. Then enter the prediction and analysis stage. A people-oriented algorithm should be used at this stage to reduce bias. Next is continuously evaluate the police operation rather than respond immediately. The most important thing is that the purpose of police forecasting is not to predict where the criminals really are, but to deploy the police force efficiently.

3.3 Governance approaches

The government plays a key role in addressing some moral and legal barriers. The legislature should enact laws and conduct external audit through predictive policing algorithm to find the defects of the internal system of the police (Scanlan, 2019). Moreover, the external audit will not be affected by the internal pressure of the police department.

Second, in the process of data mining, the personnel who have the opportunity to really contact the source data of crime are divided into two categories. The first category is the staff of government departments and police departments engaged in crime prediction, and the second category is the staff of computer software companies who build a data mining platform for predictive policing and develop policing software. For the first category of government workers, the government should take measures against the threat of using data mining technology to citizens' privacy and ensure the legitimacy of data use and confirm the qualification of staff who have access to police data. For the employees of the second type of computer companies, it is suggested that the government establish laws through the legislature to punish the improper manipulation of criminal data.

3.4 Educational approaches

The police believe that chaos and lawbreakers are factors threatening society. They believe that crime should take the initiative and pursue stimulation. Therefore, under the control of police culture, the police should be made aware that prejudice will cause inaccuracy in law enforcement. They should be made to improve their stubborn views on the bottom of society or marginalized groups and be impartial in the process of law enforcement. At the same time, we should train the police's data ability, strengthen the understanding of prediction models, and minimize the inflow of many invalid or discriminatory data into the source data set. To avoid accountability problems caused by lack of transparency and understanding of forecasting models.

In addition, the main source of police bias against law enforcement with criminal records are that such people are more likely to commit crimes again. The reasons for this argument are, on the one hand, the bias of the police in law enforcement, and, on the other hand, the improper placement of criminals. Criminals in prison usually focus on punitive education, while ignoring the resocialization education of criminals.

Therefore, to eliminate the prejudice of the police, it is necessary to properly resocialize the prison personnel and reduce the incentives for recidivism. In addition, strengthening the police's understanding of data mining algorithms is also a key point of view.

3.5 Other ideas

The criticism of predictive policing is usually divided into three categories. The first criticism is that police managers or software companies may abuse their power. The second is against the algorithm for predictive policing and the data set for training. The third is the possible technical and moral social impact of predictive police research.

Another idea is that people need to improve transparency during development and implementation of predictive policing. The police department needs to tell public which predictive system they used and how they assess the system. A supervisory committee also could be built to supervise the predictive policing algorithms and widespread policing activities.

Overall, although there are many moral and social challenges in police prediction, these challenges can be overcome. The third part emphasizes how to deal with the challenges according to the solutions of technology, procedure, government, and education. Therefore, predictive policing is not to identify the details of criminals, but to achieve the role of prevention.

References

- ACS Code of Professional Conduct. Acs.org.au. (2014). Retrieved 12 March 2022, from https://www.acs.org.au/content/dam/acs/acs-documents/ACS%20Code-of-Professional-Conduct_v2.1.pdf.
- Borowik, G., Wawrzyniak, Z. M., & Cichosz, P. (2018). *Time series analysis for crime forecasting*. Paper presented at the 1-10. <https://doi.org/10.1109/ICSENG.2018.8638179>
- Du, M., Liu, N., & Hu, X. (2019). Techniques for interpretable machine learning. *Communications Of The ACM*, 63(1), 68-77. <https://doi.org/10.1145/3359786>
- Guide to data analytics and the Australian Privacy Principles. oaic.gov.au. (2018). Retrieved 12 March 2022, from <https://www.oaic.gov.au/privacy/guidance-and-advice/guide-to-data-analytics-and-the-australian-privacy-principles>.
- Heaven, W. (2020). *Predictive policing algorithms are racist. They need to be dismantled.*. MIT Technology Review. Retrieved 11 March 2022, from <https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/>.
- Hustedt, C. (2016). *A Public Value Perspective on Predictive Policing in the US*. ResearchGate. <https://doi.org/10.13140/RG.2.2.14410.62404>
- Lecture Notes. (2022). Retrieved 9 March 2022, from <https://wattlecourses.anu.edu.au/mod/book/view.php?id=2471962>.
- McCue, C. (2006). *Data mining and predictive analysis: Intelligence gathering and crime analysis*. Elsevier Science & Technology.
- Momsen, C., & Rennert, C. (2022). *Big Data-Based Predictive Policing and the Changing Nature of Criminal Justice – Consequences of the extended Use of Big Data, Algorithms and AI in the Area of Criminal Law Enforcement – – KriPoZ*. Kripoz.de. Retrieved 11 March 2022, from <https://kripoz.de/2020/05/19/big-data-based-predictive-policing-and-the-changing-nature-of-criminal-justice-consequences-of-the-extended-use-of-big-data-algorithms-and-ai-in-the-area-of-criminal-law-enforcement/>.
- Nine, A. (2022). *Predictive Policing Software Shown to Entrench Bias, not Address It - ExtremeTech*. ExtremeTech. Retrieved 12 March 2022, from <https://www.extremetech.com/extreme/329476-predictive-policing-software-shown-to-entrench-bias-not-address-it>.
- Scanlan, J. (2019). Auditing Predictive Policing. *Brigham Young University Prelaw Review*, 33(4). Retrieved 13 March 2022, from <https://scholarsarchive.byu.edu/byuplr/vol33/iss1/4>.
- Spiekermann, S. (2012). The challenges of privacy by design. *Communications of the ACM*, 55(7), 38-40. <https://doi.org/10.1145/2209249.2209263>

Statement on Algorithmic Transparency and Accountability. Acm.org. (2017). Retrieved 12 March 2022, from https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf.

Yen, CP., Hung, TW. Achieving Equity with Predictive Policing Algorithms: A Social Safety Net Perspective. *Sci Eng Ethics* **27**, 36 (2021). <https://doi-org.virtual.anu.edu.au/10.1007/s11948-021-00312-x>

Appendix

Figure 1

$$s(t) = \sum_{n=1}^N \left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right) \right).$$

<https://doi.org/10.1109/ICSENG.2018.8638179>

Note: suppose P is the regular period in the time series (for example, when the time variable be scaled in days, annual data $p = 365.25$ or weekly data $p = 7$). Smooth seasonal effects will be calculated with the following method.

Figure 2

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t.$$

<https://doi.org/10.1109/ICSENG.2018.8638179>

Note: function $g(t)$ represents the trend of simulating aperiodic changes, function $s(t)$ represents seasonality in week or year, and function $h(t)$ represents the impact of holidays. The error ε_t represents any special changes that the model cannot adapt to.