Data challenge 2 JBG050

Handbook

Bennett Kleinberg

Last update: 21 April, 2023

Welcome to this year's Data Challenge 2 course (JBG050)!

This handbook provides the details necessary to start this course, the structure and timeline of the course, and the grading criteria with which we will assess your work.

The course builds on your previous data challenges and the courses you took as part of your BSc in Data Science. A key difference to previous data challenge courses is that you are now working on a real problem provided by stakeholders, which implies that you must first translate the stakeholder's problem into a *data-science problem*.

The problem this year is designed in collaboration with the Metropolitan Police in London and you will have the chance to ask questions to the key decision-makers within that police force who are involved in this project.

1 Problem description

The topic for your project occupies law enforcement agencies and - at a wider level - national policy-making in many countries, including the United Kingdom. The increasing dynamics that the police face (e.g., new forms of crime, ebb-and-flow of crime volume, new functions of policing) and cuts to police funding in previous years created a key challenge for senior managers and policymakers: how to best allocate resources to the police? In the simplest sense, resource allocation can be understood as determining the number of needed officers or the demand on a police force.[1]

Your task as a group is to contribute to this issue with the help of data science techniques to develop an automated, data-driven police demand forecasting system that can aid key decision-makers in allocating the needed resources at the right place and the right time. At the same time, we want you to think about the ethical implications that your decisions could have when applied in real-life.

You are asked to use (and free to obtain) historical crime data that contains all police-recorded crimes in Egnland and Wales since 2010[2]. These data are the starting point for your analysis.

The specific overarching question for your project is: **How can we best estimate police demand in an automated manner for residential burglary in the London borough of Barnet[3] to inform the most effective use of police resources to reduce burglary?**

In the introductory plenary session, you have the chance to ask questions to the teaching team and to the stakeholder (The Metropolitan Police London) after having read through the current document.

[1] For a more nuanced definition, see *Figure 3. Visualisation of police demand and its drivers* in Laufs et al. (2021).

[2] These are open data that anyone can obtain.

[3] https://en.wikipedia.org/wiki/London_Borough_of_Barnet

1.1 Background information

Burglary broadly is typically tackled in three ways (omitting covert proactive work using sensitive capabilities)

1. Getting to or near the scene quickly enough to catch suspects in the act.
2. After the fact, via detective work – forensics, CCTV, witness identification, etc.
3. Preventatively, by creating an environment that is undesirable to burgle, by increased police presence, target hardening and offender diversion.

All of these opportunities are made more effective by an accurate understanding of burglary patterns: knowing when, where and why burglaries may occur. This allows police to either discourage the offence by visible presence, target hardening, or creates more opportunities to catch offenders in the act.

However, in densely populated areas with high value neighbourhoods and finite resourcing, tackling the burglary problem has proved maddeningly difficult. The London borough of Barnet has one of the highest burglary rates in London and this has been a perennial problem. There are many desirable targets, a very dense population, and many other issues that need police attention and thus, this is one of London's truly wicked problems. Despite knowing all the tricks and tactics, this remains one of the central challenges for the police.

1.2 Expectations

To answer this question most usefully, your projects should provide forecasts as tightly as possible with the data, with a clear proposal of where and when the next burglaries may take place. Longer term analyses (i.e. multi-year and seasonal trends) would also be useful. For the most viable solutions, your group should have an eye on what works in crime reduction approaches (i.e. what is known from the literature). A starting point for this background is: https://www.college.police.uk/research/what-works-centre-crime-reduction

1.3 Problem resource limitations

There are resource limitations. For simplicity, we focus on what neighbourhood police teams can achieve in prevention and intervention.

The borough of Barnet is split into 24 wards. You can assume that at any given day there are 100 police officers for the whole of Barnet available to patrol between the hours of 0800 and 2200. Once assigned to a ward, these officers cannot go outside their ward boundaries. However, due to other policing priorities and demands they can only be used specifically against burglary for 2 hours on 4 days a week (i.e., 2 hours x 100 police offers = 200 hours per day).

Special operations can be planned for specific times/dates using more officers. This can be done no more than once every four months.

You can find ward boundaries via https://www.barnet.gov.uk/sites/default/files/2022-02/aLBB_WardsPDs_May22.pdf and https://www.barnet.gov.uk/elections-and-voting/barnet-ward-boundaries#title-2

1.4 The data

Your primary source of data is a dataset (obtainable from the open data archive of https://data.police.uk/) that spans all police-reported crimes since December 2010 in the whole of England, Wales and Northern Ireland (in total more than 70 million cases). Each reported crime is defined to belong to a specific crime type (e.g., violent crime, burglary) and spatio-temporal details. Importantly, to preserve the privacy of affected individuals, the data are aggregated to a monthly level (e.g., 3rd of Feb, 2015 is aggregated to Feb, 2015) and - on the spatial dimension - to a *lower-super output area* (LSOA)[4]. The LSOAs are units provided by

[4] Before that spatial aggregation is happening, each case's geospatial coordinates are slightly randomised.

the Ordnance Survey and used for census-related purposes. You can find details about that anonymisation procedure below in the suggested reading. In England, there are currently 32,844 LSOAs.

For your project, you are free to use other official and publicly available datasets that could help you provide better insights for the stakeholder. For example, the UK Office of National Statistics makes available so-called *deprivation indices* which measure societal well-being for each LSOA, and you can obtain various other datasets on LSOAs.

1.5 Accessing the data

- You can access the data for download from the police.uk data repository via this link: https://data.police.uk/data/
- In addition to crime (and anti-social behaviour) incident data, the police.uk data archive also provides you with the case outcomes (e.g. whether a crime incident went to court) and stop-and-search data. You can - but do not have to - incorporate and/or use that data for your project.
- Data on spatial boundaries of police forces and neighbourhoods: https://data.police.uk/data/boundaries/
- Additional data that could be useful is the UK's societal well-being data measured through the "index of multiple deprivations" (IMD). You can find these data here: https://opendatacommunities.org/def/concept/folders/themes/societal-wellbeing. More information on the IMD can be found at https://www.gov.uk/guidance/english-indices-of-deprivation-2019-mapping-resources#indices-of-deprivation-2019-explorer-postcode-mapper

You are free to use any other official, publicly available data for your project.

1.6 Background reading

- The anonymisation procedure of the police.uk data: https://data.police.uk/about/#anonymisation
- Tompson et al. (2014). UK open source crime data: accuracy and possibilities for research. *Cartography and Geographic Information Systems*. https://www.tandfonline.com/doi/full/10.1080/15230406.2014.972456
- Laufs et al. (2021). Understanding the concept of 'demand' in policing: a scoping review and resulting implications for demand management. *Policing and Security*. https://www.tandfonline.com/doi/full/10.1080/10439463.2020.1791862

Reference guide on forecasting:

- Hyndman & Athanasopoulos (2018). Forecasting: Principles & Practice. Freely available at https://otexts.com/fpp2/[5]

1.7 Software/resources for this project

You are free to use any software that you have access to. As a group, you will have to share your documented code with us via GitHub at the end of this course.

2 London trip

Last year, we travelled with the two best groups to London and presented the findings to an audience of decision-makers from a policy, operational and strategic level of the Metropolitan Police at their headquarters in New Scotland Yard.

[5] This book is R-based, but the principles and underlying statistical foundations are useful even if you work in python

We will have a similar trip to London this year again. The universities (TiU and TU/e) will pay for for the trip (travel, accommodation and subsistence) and the two winning groups will be able to fine-tune their presentation skills in a workshop before the London trip.

3 Structure of this course

3.1 Differences to previous "Data Challenges"

The fundamental objective of this course is to take you a step further to a real-world Data Science project with all the issues and uncertainties it brings.

In prior Data Science courses, including the other Data Challenges, you have been learning Data Science in an academic environment as follows: I am given problem X to solve, so I have to find the solution for X.

Data Science practice is nothing like that at all.

In practice, stakeholders state X as the final objective to solve, but along the way, you figure out that before solving X, you first have to solve Y and Z, and then time runs out. This makes real-life data science projects messy and introduces uncertainty. The outcome is often not reaching the objective but moves you closer to a solution. And that is all everyone expects.

We designed this course, "Data Challenge 2", to take you a step further to real-world Data Science in practice in a safe environment by giving you a real-life case but providing you with supervision and guidance along the way.

The objectives for this course are that you learn to:

- translate a stakeholder's problem into a Data Science problem
- specify a sub-problem from a broader problem
- apply Data Science techniques to address that sub-problem
- gradually uncover the issues that have to be addressed before "solving" the problem
- refine your project to make a meaningful step towards X
- handle and resolve uncertainty (experiencing that uncertainty is a necessary part of it)

It is important to re-iterate the following: not all problems have a clear solution (no one expects this from you in messy real-world data science).

Thus, while the assignment makes you experience uncertainty, the course is exactly about you going through it. We specifically designed the course to give you time to make mistakes and to help you learn from them.

The tutors and lecturers for this course are there to help you along the way.

3.2 Lecturers

- Dr Bennett Kleinberg, Assistant Professor in Data Science, Department of Methodology and Statistics, Tilburg University (responsible lecturer and coordinator)
- Dr Laura Genga, Assistant Professor in Information Systems, Department of Industrial Engineering and Innovation Sciences, TU/e

3.3 Tutors

Below you can find a list of the tutors.

- Paula Dodig p.dodig@student.tue.nl
- Alexander Liu a.liu1@student.tue.nl

- Paul Michielsen p.j.c.michielsen@student.tue.nl
- Gaby Does g.m.does@student.tue.nl
- Srinidhi Ilango s.srinidhi.ilango@student.tue.nl
- Cameron Dougherty c.c.dougherty@student.tue.nl
- Tisha Amara Letitia Saragi a.l.saragi@student.tue.nl
- Rohan Babani r.n.babani@student.tue.nl
- Andra Arandra Kalista Malkan a.arandra.kalista@student.tue.nl
- Cheuk Lam Mo c.mo@student.tue.nl

3.4 Group formation

The groups have been formed by us (using random group member assignment), and you can find your group on Canvas.

3.5 Working with your group

All activities for this course run on campus (all in Eindhoven).

Your tutor will schedule weekly meetings with your group and will let you know in which room these take place. The time for group meetings is in line with the official timetable:

- Wednesdays: 13:30 - 17:30 (Eindhoven campus)
- Fridays: 8:45 - 12:45 (Eindhoven campus)

You do not have to fill the whole time with the group meetings, but all meetings with your tutor should happen in these time blocks in your rooms. You are free (and will have) to hold additional group meetings.

3.6 The meeting log

It is important that you carefully prepare your group meetings and log what has been discussed in them. To facilitate this, each group is expected to keep a meeting log (of each meeting, including the meetings with tutors, lectures and you group alone).

You can see below that part of your grade is the meeting log (i.e., you have to submit your group's meeting log, and every group member has to have played the role of log keeper at least once).

The log serves three purposes:

1. it sets an agenda of points you want to address in the meeting (keep in mind that time is of the essence for such projects)
2. it provides clarity for all group members of what has been discussed
3. it ensures that tasks are distributed, and every group member knows who is preparing what

The log for each meeting must include the following:

- General information
  - Group members present
  - Group members absent (incl. the reason)[6]
  - Date and time of the meeting
  - Location of meeting
  - Name of log keeper[7]
- Meeting content

[6] Note that it is expected that you attend every group meeting. If there are valid reasons why you cannot attend a meeting, you have to discuss this with your group tutor in advance and obtain their approval.

[7] Note: every group member is expected to prepare the log for at least one meeting

- Agenda of the meeting
- Summary of points discussed (this can be in bullet points)
- Tasks for the next meeting (who will do what?)

You need to submit the meeting logs at the end of the course. *It is important that you log each meeting directly after it happened.* Make sure to also share the log with your group and tutor.

Note: the meeting log keeper role is not the same as the SCRUM master role.

4 Timelines

4.1 General schedule

We will have four plenary sessions. In your group, you will hold two group meetings every week, one of which is with your group's tutor.

Week	Day	Date	Activity
1	Wed	26/04/2023	Plenary session 1
1	Fri	28/04/2023	UNI CLOSED
2	Wed	03/05/2023	Group work
2	Fri	05/05/2023	Group work
3	Wed	10/05/2023	Group work
3	Fri	12/05/2023	Group work
4	Wed	17/05/2023	Plenary session 2
4	Fri	19/05/2023	Group work
5	Wed	24/05/2023	Group work
5	Fri	26/05/2023	Group work
6	Wed	31/05/2023	Plenary session 3
6	Fri	02/06/2023	Group work
7	Wed	07/06/2023	Group work
7	Fri	09/06/2023	Presentations
8	Wed	14/06/2023	Group work
8	Fri	16/06/2023	Group work
9	Wed	21/06/2023	Plenary session 4
9	Fri	23/06/2023	Group work

4.2 Milestones

The timeline below is a suggestion of milestones. These are not harsh deadlines but give you a way to better navigate the problem space as a group.

Week	Milestones for this week
1	read in the data / understand the data / clean the data / pitch first thoughts
2	determine the unit of analysis / aggregate and subset the data / think about the specific approach you want to take
3	decide on final dataset / gather additional data if needed / prepare data for your analysis / pitch analysis ideas
4	run the analysis / use the data to answer your sub-questions
5	refine the the analysis
6	draw conclusions from the data / determine how to evaluate your approach / fine-tune analysis / prepare presentation

Week	Milestones for this week
7	finalise the presentation; presentation sessionfinalise forecasting model(s) / work in presentation and technical report
8	work on the technical report and the ethics reflection
9	finalise all deliverables

5 Grading

5.1 Deliverables and deadlines

Four deliverables count towards your final grade. You can find details on how each component is weighted in the grading criteria section.

1. The group presentation (deadline: 8 June 2023, 23:59h) - 50% of the final grade
2. The final report (deadline: 23 June 2023, 23:59h) - 20% of the final grade
3. The discussion of ethical consideration (deadline: 23 June 2023, 23:59h) - 15% of the final grade
4. General progress as a group (to be evaluated by the group tutors) - 5% of the final grade
5. The meeting log (updated continuously, submitted at the end of the course) - 5% of the final grade
6. Code documentation on GitHub - 5% of the final grade

You will receive a grade as a group. If problems arise within a group (e.g., if a group member disengages), we may opt for individualised grade deductions.

5.1.1 The group presentation

The final presentation is targeted at a stakeholder audience and should include:

- a recap of the problem and your specific sub-problem
- the aims of your project
- the decisions made with the data
- the findings of your analysis
- 3 core conclusions and 3 recommendations (derived from the conclusions)

The presentation must be no longer than 10 minutes. The speaking time should be distributed equally among the group members. The slides used for your presentation should be uploaded to Canvas by the 8th of June, 23:59h.

The presentation will be presented on campus on the 9th of June.

Tournament-style presentations:

The presentations will be used to determine which groups will travel to London to present their final work to the police at New Scotland Yard (HQ of the Metropolitan Police). That tournament may take the following form:

- Round 1: each group is part of a pool of other groups (e.g., 7 groups per pool in total) and holds a presentation (10 mins) in front of the teaching team and the pool of students
- Round 2: from each pool, the best groups are selected to present in a final in front of the whole teaching team and all students. Groups in the final will have two additional minutes to present their work (12 minutes).
- Decision: the two best groups from the second round will travel to London (the trip is likely to happen in the week of 19 June 2023)

5.1.2 The final report

The final report (written as a group) should cover:

- a brief introduction to the problem
- details on your data science approaches
- a detailed, statistical evaluation of your approach(es)
- a link to the GitHub page where your code and data are located and documented[8]
- an interpretation of your finding(s)
- at least three limitations of your approach and suggestions (for each limitation) how these could be fixed in the future

Using the provided template, the report should not be longer than **six** pages (including references and the reference list). The report must be submitted before the deadline as a pdf file named "group_X_technicalreport.pdf" (where X is replaced with your group number).

Template for the technical report

You should use the following Overleaf template: https://www.overleaf.com/latex/templates/acl-rolling-review-template/jxbhdzhmcpdm - taken from the proceedings of the Association of Computational Linguistics conference (ACL). When submitting your final report, make sure to use the non-anonymised version and list all group members as authors. The permitted page length includes the in-text references + the reference list. Your submission should be as a pdf via Canvas

Guide to GitHub:

In case you are unfamiliar with GitHub, have a look at these tutorials/guides:

- GitHub Quickstart Hello World
- GitHub Tutorial - Beginner's Training Guide
- GitHub desktop GUI

You can (but do not have to) use GitHub for your whole data and code flow. All we require is that you share a link to a public GitHub repository with the code to run your analyses and the data used.

5.1.3 Discussion of ethical considerations

A big challenge for real-world data science projects is the balance between stakeholder (in the widest sense) interests and potential ethical problems that may arise either directly or indirectly as a consequence of the project outcomes. In this course, you are working an a particularly challenging topic around police data, which has long been a controversy, but has also resulted in several safe-guarding frameworks to ensure proper applications of automated decision-making frameworks in the area of law-enforcement.

A whole body of work - both from legal scholars as well as computational crime scientists - has critically examined the role of predictive approaches in policing. Some core papers (incl. rare empirical evidence), are listed hereafter:

- Albert Meijer & Martijn Wessels (2019) Predictive Policing: Review of Benefits and Drawbacks, International Journal of Public Administration, 42:12, 1031-1039, DOI: 10.1080/01900692.2019.1575664
- Lyria Bennett Moses & Janet Chan (2018) Algorithmic prediction in policing: assumptions, evaluation, and accountability, Policing and Society, 28:7, 806-822, DOI: 10.1080/10439463.2016.1253695
- P. Jeffrey Brantingham, Matthew Valasik & George O. Mohler (2018) Does Predictive Policing Lead to Biased Arrests? Results From a Randomized Controlled Trial, Statistics and Public Policy, 5:1, 1-6, DOI: 10.1080/2330443X.2018.1438940
- Susser, Daniel, Predictive Policing and the Ethics of Preemption (June 29, 2021). Ben Jones and Eduardo Mendieta (eds.), The Ethics of Policing: New Perspectives on Law Enforcement (NYU Press), 2021, Available at SSRN: https://ssrn.com/abstract=3875917

[8] The code should enable someone else to fully replicate your findings

One of the most influential frameworks to think about ethics in data applications in UK policing is the ALGO-CARE framework (Oswald et al., 2018) that is now common practice and required for all police data science projects in the UK.[9]

Your task for the ethical considerations report is as follows:

- use the ALGO-CARE framework[10] on your group's project to *identify*, *discuss*, and *make suggestions for improvements* of potential ethical problems
- choose 3 different ALGO-CARE criteria and for each of them:
  - describe how each comes back in your project work (= identifying the problem)
  - discuss the implications of each ethical reservation/problem (e.g., who is affected, how are they affected?)
  - provide a suggestion for mitigating each ethical concern (i.e. what could be done to remove your concerns?) and mention the impact that the decision has had (or might have had) on the project outcomes
- discuss ethical considerations of technical design choices of your project (e.g., what are the implications of a specific analytical procedure?)

Further relevant literature that could help you:

- Brayne, Sarah, Alex Rosenblat, and Danah Boyd. "Predictive Policing". Data & Civil Rights: a new era of policing and justice. October 27, 2015. https://datacivilrights.org/pubs/2015-1027/Predictive_Policing.pdf
- B. Green, "Data Science as Political Action: Grounding Data Science in a Politics of Justice," in Journal of Social Computing, vol. 2, no. 3, pp. 249-265, September 2021, doi: 10.23919/JSC.2021.0029. https://ieeexplore.ieee.org/abstract/document/9684742

Page limit: 3 pages. A template is provided below. The report must be submitted before the deadline as a pdf file named "group_X_ethics.pdf" (where X is replaced with your group number).

Template ethics report

You should use the following Overleaf template: https://www.overleaf.com/latex/templates/acl-rolling-review-template/jxbhdzhmcpdm - taken from the proceedings of the Association of Computational Linguistics conference (ACL). When submitting your final report, make sure to use the non-anonymised version and list all group members as authors. The permitted page length includes the in-text references + the reference list. Your submission should be as a pdf via Canvas

5.1.4 The meeting log

See the details above –> The meeting log. The meeting log will need to be submitted via Canvas.

5.1.5 General progress as a group

In close coordination with your weekly supervisors, each group will receive a lightweight grade for general progress made throughout the course.

5.2 Grading criteria

These are updated at the moment.

[9] If you are interested in a more extensive review of "data analytics and algorithms in policing", you can have a look at this article from RUSI or the briefing version of this report here.

[10] Marion Oswald, Jamie Grace, Sheena Urwin & Geoffrey C. Barnes (2018) Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality, Information & Communications Technology Law, 27:2, 223-250, DOI: 10.1080/13600834.2018.1458455