Applied Machine Learning: Tutorial Number 4 Answers

School of Informatics, University of Edinburgh
Instructors: Oisin Mac Aodha and Siddharth N.

September 2022

In this tutorial we will explore ethical challenges in data science and machine learning by focusing on two case studies. Read the case studies at the end of this document before the tutorial session and attempt to answer the questions listed below. Make notes for yourself to use during the session.

During the tutorial we will have a discussion in smaller groups, for at least one of the case studies. You do not have to reach a unanimous conclusion. If there is disagreement, make a note of it, and try to work out why you are disagreeing – you may be working from different assumptions.

The case studies below have been adapted from 'An Introduction to Data Ethics' by Shannon Vallor. This workshop is heavily inspired by Foundations of Data Science Workshop 1 by David Sterratt and Kobi Gal.

Questions

(a) What are the most relevant ethical challenges for data scientists and machine learning practitioners that are present?

(b) List the various stakeholders involved. What was at stake for each of them? Try to include all possible affected stakeholders.

(c) What specific significant harms to members of the public are relevant here? List as many types of harm as you can think of.

(d) How should those potential harms have been evaluated alongside the prospective benefits? Could the benefits hoped for have been significant enough to justify the risks of harm you identified above?

(e) How did the individuals responsible defend their position, and how ethically valid is this justification?

(f) Describe several things that the individuals responsible could have done differently, to acquire the intended benefits in a less harmful way.

Solution:

Case Study 1: Loan Evaluation

(a) Appropriate data collection and use: It does not appear that the data used to train the system has been validated for quality or accuracy.

Personal, social, and business impacts: The potential impact of systematically biased loan decisions has not been considered. There could be biases in decisions related to the gender, race, or ethnicity of the loan applicants.

Data storage, security and responsible data stewardship: There are questions related to the indiscriminate collection and sharing to third-parties of personal data, potentially without explicit consent being granted.

Human accountability in data practices and systems: The loan officer is responsible for delivering the decision but puts the blame on the automated system.

(b) The loan applicants (Fred and Tamara), who wanted a loan, but who risked getting a refusal and potentially impacting their credit score

The bank loan officer, whose potential over reliance on automated systems risks replacing their job

The bank, whose reputation and therefore profits were at stake if it turns out that their decisions are systematically biased

The software developers, who risk potential repercussions from having non-transparent systems or questionable data procurement practices

(c) There are harms related to the lack of transparency. In this context, transparency is the ability to see how a given social system or institution works, and to be able to inquire about the basis of life-affecting decisions made within that system or institution. Here transparency will be served by individuals having access to information about exactly why they were denied a loan, and by whom.

There are also harms related to fairness and justice. The most common causes of such harms are: arbitrariness; avoidable errors and inaccuracies; and unjust and often hidden biases in datasets and data practices. Here harms could be driven by poor quality, mislabeled, or error-riddled data (i.e., 'garbage in, garbage out'); inadequate design and testing of data analytics; or a lack of careful training and auditing to ensure the correct implementation and use of the data system.

(d) The potential benefits is having an automated system that could potentially be free from any biases or profiling that human loan officers may inadvertently use. The bank should have done more work in identifying any sources of bias in the automated system and validate if it is better (i.e. causes less harm) than humans. One potential way to achieve this would be to validate the system on data from customers that have previously been given loans by the bank as they will have ground truth knowledge indicating if they paid back their loan on time or not.

(e) The loan officer simply stated that the system said the applicants were at a high risk of defaulting on their loan. They did not provide any justification of this decision and it seems that, even if they wanted to, they would have be unable to as the system does not provide this type of information. This seems like a weak justification as ultimately their employer is responsible for choosing this system in the first place.

(f) The loan officer: could have insisted that they bank used a transparent system for evaluating loan applications.

The software developers: could have provided more insight into what the model was trained on, provide greater transparency related to any biases that may have been identified during their model development and validation, and could have chosen a machine learning algorithm that is transparent by design.

Case Study 2: OK Cupid

(a) Appropriate data collection and use: The data was collected (scraped) in a way that OK Cupid had not intended it to be and which was not in line with the terms and conditions OK Cupid users had agreed to.

Personal, social, and business impacts: The effect of the data release on users was not considered sufficiently; users' life interests, autonomy, dignity, privacy and relationships could have been damaged.

Data storage, security and responsible data stewardship: the privacy protection was not adequate.

Human accountability in data practices and systems: who was accountable, the researchers or the university for which they were working? The University put the blame squarely on the researchers.

(b) The researchers, who wanted a good paper, but who risked losing their reputation.

The OK Cupid users, whose privacy was at stake.

The OK Cupid company, whose reputation and therefore profits were at stake.

Aarhus University, whose reputation for ethical scientific research was at stake.

The journal and reviewers, who risked their reputation publishing work that has not been undergone appropriate internal ethical review.

(c) Privacy: intimate details of OK Cupid's users' lives were made public, given that there was not sufficient anonymisation.

Autonomy: users whose data were exposed would have compromised their privacy.

(d) There should have been a proper ethics procedure undertaken, meeting standards such as those laid down in GDPR (though GDPR was not in place at the time). Consent would have needed be obtain from any individual participating in the study.

(e) The data was already public - but the data was not public in the sense that any internet user could see it without an account; an account was required which had terms and conditions attached that prevented the use of the data.

That there was no legal problem – This appears not to be true (as OK Cupid were able to force the data to be taken down) but this does not constitute an evaluation of the potential ethical harms and benefits and does not imply no harm was done.

"Don't know, don't ask" - ignorance is not a defence under the law.

(f) They could have designed a different study that did not involve personal data on a dating website to answer the same question. For example, recruiting members of the online public and asking them to participate in an online study about relationships with e-dating sites.

If they had really wanted to use this dataset, they should have first asked OK Cupid about the possibility and come to an agreement with OK that allowed for the informed consent of the data subjects.

Behaved more ethically and responsibly once accused.

The paper reviewers could have requested more information to check if the paper had passed appropriate internal ethical review at the host institution. They could have also asked the researchers to confirm if the data they scraped was within the terms of use of the website.

The journal could have had a more robust process in place to ensure that data with potentially individually identifying information is not disclosed.

Case Study 1: Loan Evaluation

Fred and Tamara, a married couple in their 30's, are applying for a business loan to help them realize their long-held dream of owning and operating their own restaurant. Fred is a highly promising graduate of a prestigious culinary school, and Tamara is an accomplished accountant. They share a strong entrepreneurial desire to be 'their own bosses' and to bring something new and wonderful to their local culinary scene; outside consultants have reviewed their business plan and assured them that they have a very promising and creative restaurant concept and the skills needed to implement it successfully. The consultants tell them they should have no problem getting a loan to get the business off the ground.

For evaluating loan applications, Fred and Tamara's local bank loan officer relies on a machine learning-based software package that makes use of data purchased from hundreds of private data brokers. As a result, it has access to information about Fred and Tamara's lives that goes well beyond what they were asked to disclose on their loan application. Some of this information is clearly relevant to the application, such as their on-time bill payment history. But a lot of the data used by the system as features is of the sort that no human loan officer would normally think to look at, or have access to – including inferences from their drugstore purchases about their likely medical histories, information from online genetic registries about health risk factors in their extended families, data about the books they read and the movies they watch, and inferences about their racial background. Much of the information is accurate, but some of it is not.

A few days after they apply, Fred and Tamara get a call from the loan officer saying their loan was not approved. When they ask why, they are told simply that the loan prediction system rated them as 'moderate-to-high risk.' When they ask for more information, the loan officer says he doesn't have any, and that the software company that built their loan system will not reveal any specifics about the proprietary algorithm or the data sources it draws from, or whether that data was even validated. Fred and Tamara ask if they can appeal the decision, but they are told that there is no means of appeal, since the system will simply process their application again using the same algorithm and data, and will reach the same result.

Case Study 2: OK Cupid

In 2016, two Danish social science researchers used data scraping software developed by a third collaborator to amass and analyze a trove of public user data from approximately 68,000 user profiles on the online dating website OkCupid. The purported aim of the study was to analyze "the relationship of cognitive ability to religious beliefs and political interest/participation" among the users of the site.

However, when the researchers published their study in the open access online journal Open Differential Psychology, they included their entire dataset, without use of any anonymizing or other privacy-preserving techniques to obscure the sensitive data. Even though the real names and photographs of the site's users were not included in the dataset, the publication of usernames, bios, age, gender, sexual orientation, religion, personality traits, interests, and answers to popular dating survey questions was immediately recognized by other researchers as an acute privacy threat, since this sort of data is easily re-identifiable when combined with other publicly available datasets.

That is, the real-world identities of many of the users, even when not reflected in their chosen usernames, could easily be uncovered and relinked to the highly sensitive data in their profiles, using commonly available re-identification techniques. The responses to the survey questions were especially sensitive, since they often included information about users' sexual habits and desires, history of relationship fidelity and drug use, political views, and other extremely personal information. Notably, this information was public only to others logged onto the site as a user who had answered the same survey questions; that is, users expected that the only people who could see their answers would be other users of OkCupid seeking a relationship. The researchers, of course, had logged on to the site and answered the survey

questions for an entirely different purpose—to gain access to the answers that thousands of others had given.

When immediately challenged upon release of the data and asked via social media if they had made any efforts to anonymize the dataset prior to publication, the lead study author Emil Kirkegaard responded on Twitter as follows: "No. Data is already public." In follow-up media interviews later, he said: "We thought this was an obvious case of public data scraping so that it would not be a legal problem."

When asked if the site had given permission, Kirkegaard replied by tweeting "Don't know, don't ask. :)"

A spokesperson for OkCupid, which the researchers had not asked for permission to scrape the site using automated software, later stated that the researchers had violated their Terms of Service and had been sent a take-down notice instructing them to remove the public dataset. The researchers eventually complied, but not before the dataset had already been accessible for two days.

Critics of the researchers argued that even if the information had been legally obtained, it was also a flagrant ethical violation of many professional norms of research ethics (including informed consent from data subjects, who never gave permission for their profiles to be used or published by the researchers). Aarhus University, where the lead researcher was a student, distanced itself from the study saying that it was an independent activity of the student and not funded by Aarhus, and that "We are sure that [Kirkegaard] has not learned his methods and ethical standards of research at our university, and he is clearly not representative of the about 38,000 students at AU."

The authors did appear to anticipate that their actions might be ethically controversial. In the draft paper, which was later removed from publication, the authors wrote that "Some may object to the ethics of gathering and releasing this data... However, all the data found in the dataset are or were already publicly available, so releasing this dataset merely presents it in a more useful form."