Practical problems in data science
and
Ethical dilemmas and conflicts of interests

Ziemowit Dworakowski
AGH University in Krakow,
Department of Robotics and Mechatronics

1

- Some terms are understand differently by industry and DS.
- Majority of industry practitioners **have no idea** which tasks are easy and which are hard
- Majority of people view ML systems as a „magic black box" that can do anything if the operator is good enough
- Majority of industry practitioners **misjudge quality of available data**
- Majority of industry practitioners **misjudge importance of overfitting and data leakage risks**
- Majority of industry practitioners **underestimate importance of a preprocessing of data**
- Majority of industry practitioners **overestimate importance of a classification method used**
- In industry there is a high pressure towards „marketing effect" and „selling dreams" instead of real quality

2

- Majority of ML experts do not understand how difficult data acquisition could be
- Majority of ML experts are trained only on „easy" and „clean" data, with no outliers or mislabeled samples.
- ML experts like to take it the easy way, feeding vectors of data to a decision system and demand more data if it does not converge to a good result, instead of picking **better features** and use **context knowledge**
- Majority of ML experts underestimate context knowledge and overestimates the classifier capabilities
- Some ML experts poorly estimate importance of false positive indications
- Some ML experts poorly identify factors contributing to the final result

3

Required mindset

Be prepared to work with poorly prepared data
(including mislabeled samples, outliers, lack of description etc.)

Be prepared to compromise and think non-standard
(Some problems can't be solved using typical approaches and good practices)

You need a credit of trust
(can be gained based on experience and communication skills)

Protect salesmen from themselves
(„Would you like a merit check on your leaflet?")

Think about the actual goal of your work
(It is rarely „to build a classifier")

4

Required skillset

Basic engineering knowledge
to understand context and meaning of the task at hand

Good understanding of **data leakage and overfitting** problems
these are usually the most important challenges that need to be solved in the first place

Ability to learn and find sources
(*We usually work with unique applications every time*)

Basics of AI
General overview rather than deep understanding of methods' details

Experience from several projects with emphasis on **design** and **optimization** of the solution

Python (+ its AI libraries)

5

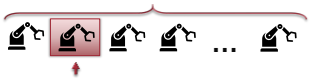
Ethical issues and conflicts of interests

6

Case study I

This is Michael
His main task is to **keep the average production cost as low as possible**

This is a population of machines under monitoring



This one particular robot is quite old, but does not show signs of significant wear yet. It can be:

A: Left as it is (**2%** risk of malfunction, **1 000 000 $** cost in such case)
B: Subjected to maintenance (**50 000 $** cost)

Michael knows that it is statistically better (lower overall cost) to leave the robot as it is, but in the unlikely case of malfunction he is afraid of the incoming discussion with the CEO…

7

Case study II

Margaret is a *data science* specialist in a software-for-surveillance company. She noticed the mistake in data acquisition routines rendering the system to be insensitive to people dressed in blue.

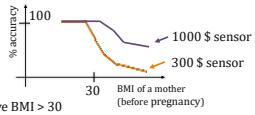
She notifies her Team Leader of the problem.
He says is too late for any changes, system goes for production.
The client will not know system's weakness.

Margaret should:
A: Listen to her team leader
B: Escalate the problem (become a „whistleblower")

8

Case study III

Helmut is a CEO in a start-up that specializes in fetus monitoring using wearable sensors. The system allows for detection of issues that, when treated early, can save life of mother and child.



Roughly 20% of potential clients have BMI > 30
A: The system should be equipped with a cheaper sensor
B: The system should be equipped with a better sensor

Values to consider:
- Easy access (*more lives saved*)
- Guaranteed success (*we can save almost every monitored life*)
- Company income (*can be used to improve system further*)

9

Case study IV

Joseph is a mechatronics' specialist. In his company he also works as a merit expert during recruitment. The company wants to hire a new employee. Joseph will be asked his opinion on the following candidates:

Rashid – great CV. Meets all the needs with ease but is „from outside", nobody knows him.

Hans – almost meets all the requirements. In addition to that, he studied with Joseph so Joseph knows that Hans learns fast and is quite nice.

He will surely quickly learn everything that is required for this position!

10

Simple heuristics to consider:

1. **Personal conflicts of interests**: disengage from the decision proces, let someone unbiased take a call.

2. **Values' conflicts:** consider a false dichotomy. Compromise or a third solution might be possible.

3. If we can't solve the conflict by ourselves, **escalate** (that is: move the conflict to a higher authority)

4. Know your **values** and learn how to **communicate effectively**

5. Know different decision making strategies, look at the big picture (*A lot of conflicts exist because people tend to overlook many possible solutions and fail to see ahead*)

11

Feedback…!

Use student surveys at the end of the semester to give feedback. Do that for as many subjects and teachers as you can. This is much more important than you think.

12