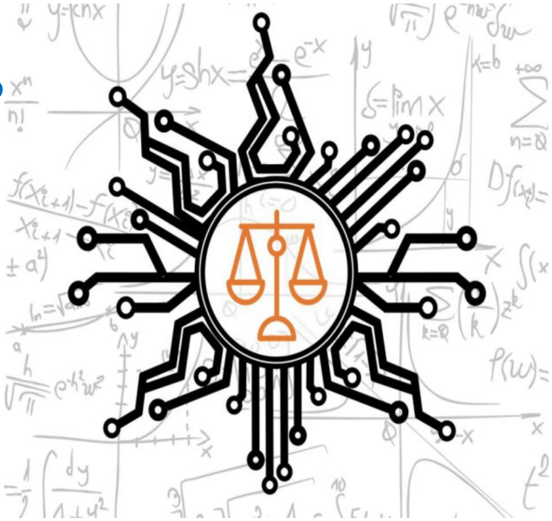
Data Ethics in Machine Learning & Data Mining

Lecture 2 (Part A)

Dr. Vassilis S. Kontogiannis
Reader in Computational Intelligence
Email: V.Kodogiannis@westminster.ac.uk
https://scholar.google.co.uk/citations?user=meTTcLAAAAAJ&hl=en&oi=ao



Ethical Issues in Data Mining (DM) and Machine Learning (ML) systems

Introduction

Data Mining (DM) and Machine Learning (ML) systems:

- can automate a lot of tedious and dangerous work now.
- are already part of our life.
- are trusted with making important decisions

But DM and ML systems:

- have innate biases which do not coincide with social norms and have no ethical grounds.
- fail in a way which is not humanly interpretable.
- can have negative economic and social impact –eliminate jobs.
- have some security issues – chat bots, autonomous cars, etc.

Challenges in DM & ML domains

In recent years, IT companies, such as **Facebook** and **Google** have:

- transformed themselves into data companies.
- built world-class AI research groups.
- accumulated a lot of Big Data about customers, not publicly available.
- made better digital marketing due to user profiling and personalization.

What's next?

- Both fields follow boom & bust cycle;
- Is society ready to accept ML/DM systems?

Some ethics definitions

- Ethics or moral philosophy

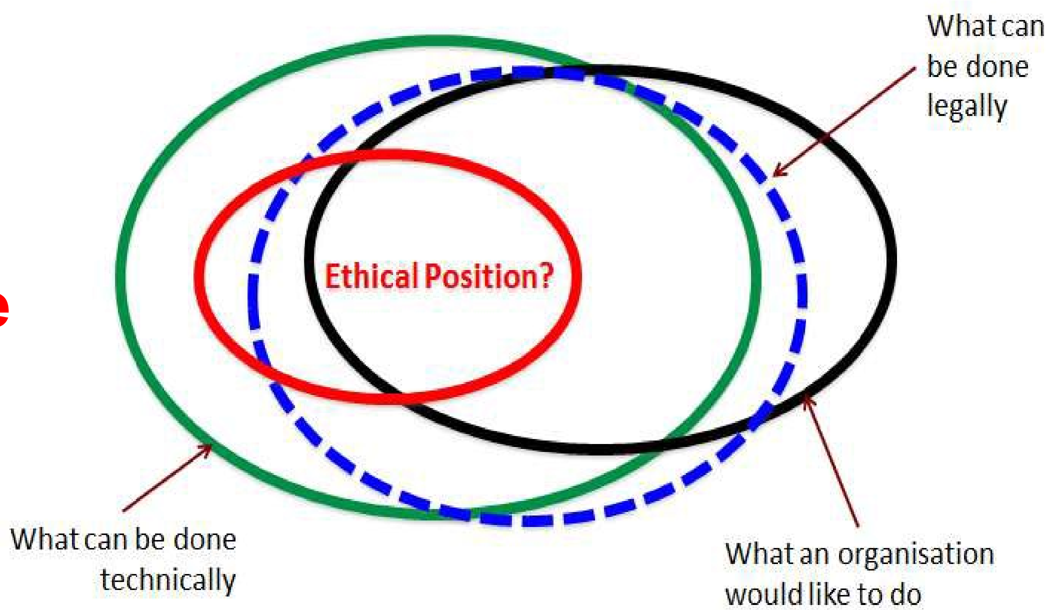
  a branch of philosophy that involves systematizing, defending, and recommending concepts of right and wrong conduct.
- Ethics vs. Laws vs. Religion

  these terms have a common root but do not coincide.
- Data ethics

  How data affects human well-being - positively and negatively.
- Ethical values

  autonomy, equality, etc.

Ethics in real life

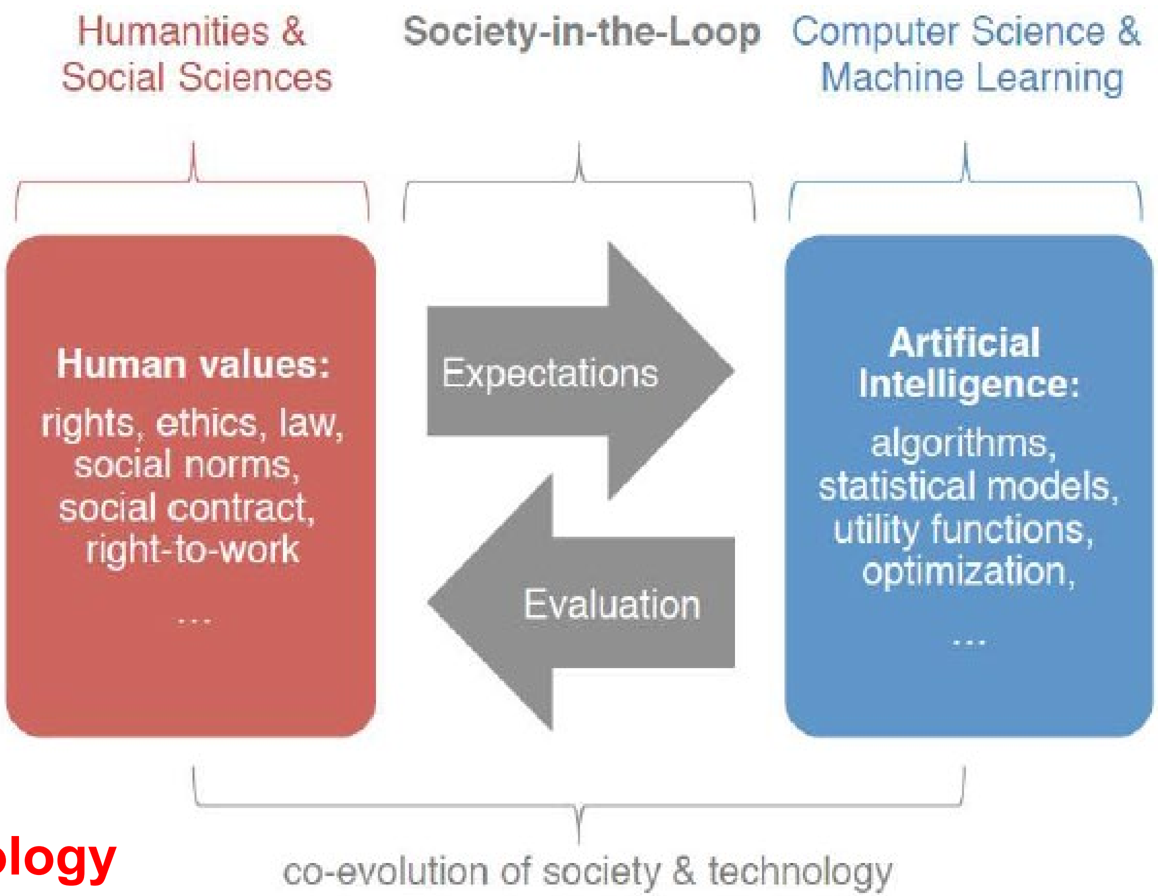


Ethics of Technology

Definition: *is an interdisciplinary research area concerned with all moral and ethical aspects of technology in society. (Luppicini,* 2008)

It views society and technology as interrelated and aims to:

- use technology ethically.
- prevent misuses.
- guide new technological advances.
- benefit society.



Ethics, Society and Technology

Current ethical DM/ML problems

- Fairness, Discrimination
- Ownership
- Transparency
- Privacy
- Accountability
- Anonymity
- Confidentiality
- Identity
- Reputation

Ethical DM/ML cases

- The Facebook emotions study (2014) - psychological research
- Panama papers (2016) - use of hacked data
- Cambridge Analytica case (2018) - psychological profiling

Ethical DM/ML cases in the near future:

- Autonomous cars
- Autonomous weapons
- meaningful human control?
- Internet of things (IoT)
- Personalized medicine (genomic information)
- Social Credit System

Ethical DM/ML issues

- Innovators are restricted to the given state of scientific and technical knowledge.
- Each technical innovation brings risks and benefits.
- How to manage risks, when implementing an innovation?

How to solve ethical issues

- What approach is best for solving DM/ML ethical issues?
  - strict national regulation vs. international regulation vs. looser code of ethics?
- Different approaches/priorities:
  - development of technology
  - businesses growth; more investments in DS/ML field
  - public interest
- Innovation first or Regulation first policy.

Legislation

- Falls behind technological progress for most DM/ML ethical concerns.
- A long tradition of regulation for consumer, security, and privacy protection in the USA.
- EU scores ahead in 2018 with GDPR.

Data privacy:

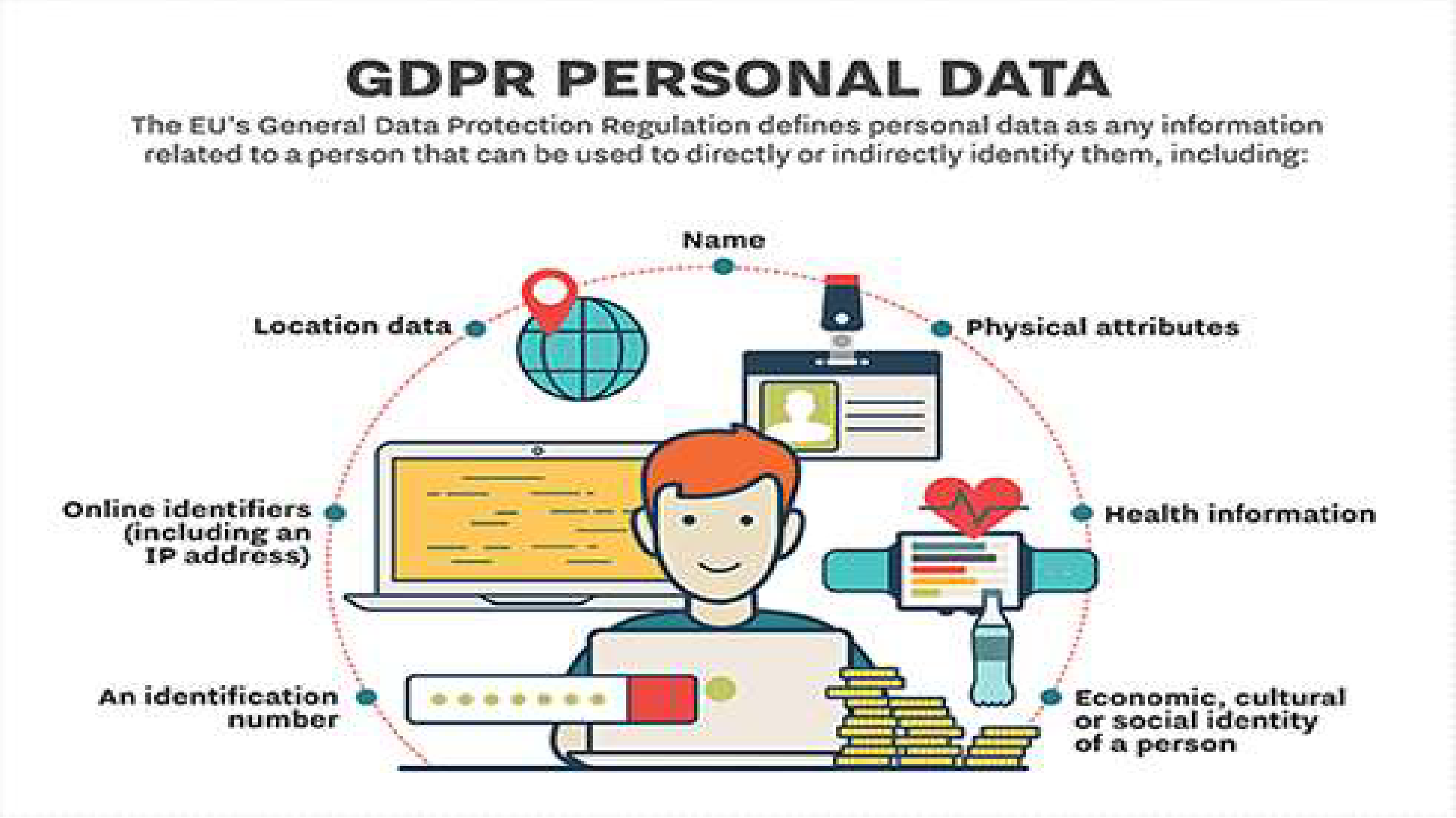
- has been already a major concern for public opinion and a political issue.
- has been already introduced into legislation.

While other DM/ML ethical issues:

- are still a subject of debate and are not fully introduced into legislation.
- there are similar issues in other fields regulated by other laws.

GDPR

Legally binding regulation, not a directive or a recommendation. Expanded definition of **personal data** – including **person's name**, **location**, **online identifiers**, **biometrics**, **genetic information**, etc.



GDPR – requirements for data protection

1. Big data analytics must be fair.
No bias and discrimination. Consumers should be awarded for data collection. Processing should be transparent.

2. Permission to process data.
Unambiguous consent from users. User consent for data use by third parties.

3. Purpose limitation.
No further processing incompatible with the original purpose.

4. Holding on data.
Using only data you need to process for a specific purpose.

5. Accuracy.
Incorrect data must be dismissed. Big data should not represent a general population. Hidden biases in data should be considered in final results. No discrimination during profiling.

6. Individual rights and access to data.
Individuals should be allowed to access their own data.

7. Security measures and risk.
Security risks should be specifically addressed during processing.

8. Accountability.
Big data processing without a defined hypothesis might cause problems. Biased profiling, too.

9. Controllers and processors.
No clear definition as both operations are performed by ML algorithms.

Current issues in ML algorithms

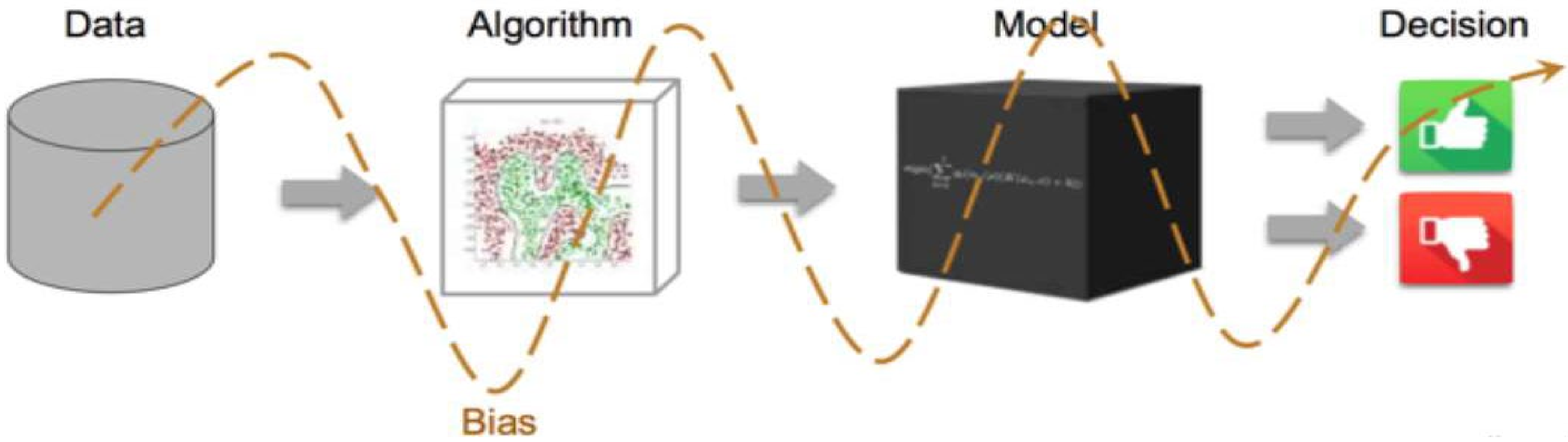
- Hard to explain the final decision to users since ML systems look like black boxes (NN-based algorithms).
- Some of the current ML algorithms behave unfair.
- DM/ML systems need to be used by professionals outside engineering/math communities.
- DM/ML systems should be incorporated into social and legal systems.

Bias issues in DM/ML algorithms

Types of biases we are interested in:

- **Algorithmic bias** (feature or model selection)
- **Data bias** (biased or irrelevant data)
- **Interpretability/Transparency** of DM/ML systems - (**model bias**)

Bias issues in DM/ML algorithms



Data bias

- is the most important component of the bias of the whole DM/ML system now; comes from data sampling.
- is a responsibility of the designer of the DM/ML system to deal with it.
- is due to various standards for datasets; no strict requirements for data content; each dataset is biased to some extent.

Sources of data bias

1. Data is a social mirror.

If training data reflects existing social biases against a minority, the algorithm is going to incorporate it.

2. The sample size disparity.

Less data available about minorities – models of minorities tend to be worse than those of the general population.

3. Cultural differences.

The statistical patterns that apply to the majority might be invalid within a minority group. A variable positively correlated with the target in the general population might be negatively correlated in a minority group – diverse names in ethnic groups.

4. Undesired complexity.

Many different overlapping minorities data groups – the combination of separate classifiers for them is complex.

5. Noise and the meaning of 5% error.

Error value can depend on the type of data or on the ML algorithm itself.

Data bias reduction

The data bias can be reduced by:

- gathering more data from different sources, thus avoiding sampling bias.
- removing variables in data associated with bias, e.g. age, sex, etc.
- talking to domain experts, where DM/ML systems will be used, in order to get more information (incorporation of external knowledge!).

Algorithmic bias

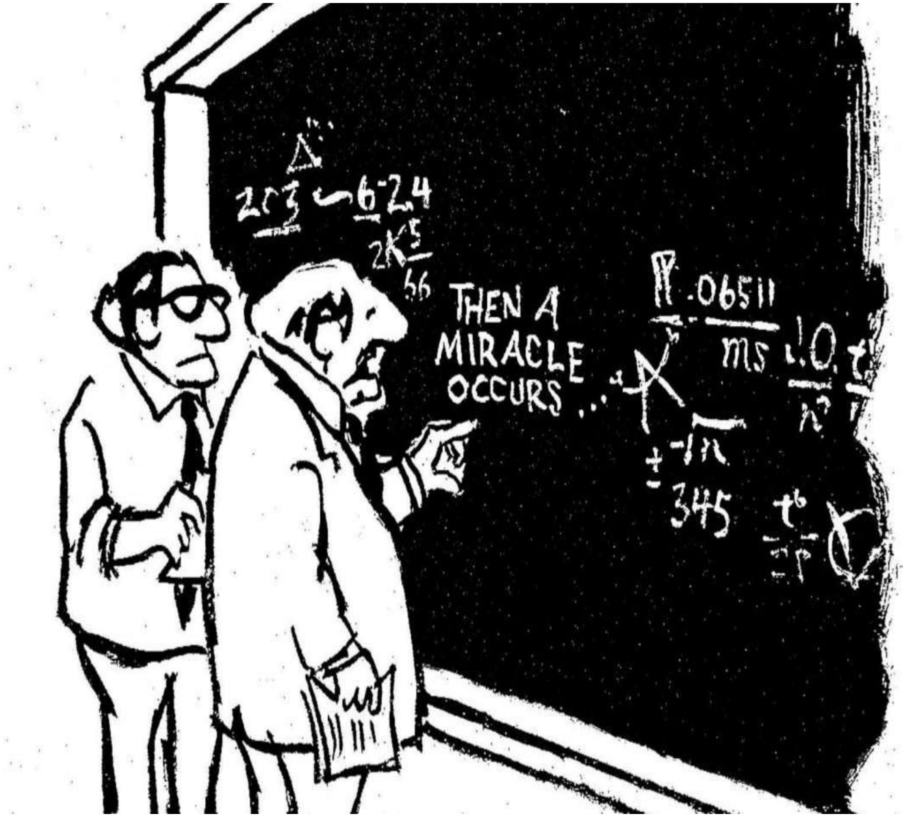
- encompasses data bias.
- leads to discrimination and unfairness.
- is introduced during the development and testing stage.
- is produced by human/programmer cognitive biases.
- has unintentional nature.

Algorithmic bias reduction

- first identify it – talk to domain experts, where DM/ML systems will be used, to get more information.
- introduce de-biasing algorithms and use de-biased datasets.
- perform external auditing and apply special regulations.
- increase algorithmic transparency.

Algorithmic **interpretability/explainability**:

- is especially important for automatic medical diagnostic software.
- relates to the legitimacy of decisions in social/business systems.
- leads to **Accuracy vs. Explainability trade-off** in various fields of applications.



DM/ML algorithm interpretability

Interpretable DM/ML algorithms:
Decision forests, linear models, Naïve Bayes, Fuzzy systems and KNN classifiers (White Box)
Non-interpretable ML algorithms:
Neural networks, SVM, kernel methods, etc. (Black Box)

Principles of accountable algorithms

1. Responsibility
Who is responsible, if users are harmed?
2. Explainability
How much of the algorithm code and data will be disclosed?
3. Accuracy
Sources of error and their effect? Worst case scenario?
4. Fairness
Potential damages to different (social) groups by your algorithms?

Application of ML algorithms

Ethically complicated cases of DM/ML algorithms:

- gender-biased results (**discrimination**)
- racist outcome – classification of black people as “gorillas” (**discrimination**, **fairness**)
- resume filtering based on age and sex in HR industries (**discrimination, fairness**)
- invisible calculation of credit score (**transparency, accountability**)
- data brokers (**confidentiality**)
- Uber taxis price forming (**transparency, fairness**)
- predictive policing (**discrimination**, **fairness**)
- personal and psychological profiling (**privacy, discrimination, confidentiality**)

What are the consequences?

What can happen, if we do not oppose biases in DM/ML systems?

- Businesses will use biased datasets for greater profits.
- DM/ML developers will apply evaluation metrics which can amplify biases – gender or race specific.
- The wide application of DM/ML algorithms will strengthen bias and polarization in society.
- Social tension and distrust to Intelligent systems and technologies will arise.

Example Application: Credit score computation

- Credit score is a numeric expression, measuring people's or company's credit-worthiness.
- Banks use it for decision-making for credit application.
- Depends on credit history.
- It indicates how dependable an individual or a company is.

Scorecard algorithm

Def: *a standard and easy to understand credit scoring algorithm. A Binary problem:*
1st class – default – a customer fails to pay install.
2nd class – a customer pays regular installments for a given time period.

It consists of:

- building and training a statistical or a ML model.
- applying the chosen model to assign a score to every credit application.

Scorecard algorithm

- Use of ML algorithms as logistic regression, random trees, boosting, neural networks, generalized additive models
- Use of Area under curve (AUC) based on ROC analysis for model evaluation, Gini coefficients
- The data should be comprehensive – allowing few missing values, and including as many data points as possible from the financial records of customers and their payment history

Current issues

- Customers with no credit history need to be set into predefined groups.
- Wide introduction of automated credit score – aims to make markets more efficient and low cost financial services but introduces algorithmic bias.
- Incomplete data can influence negatively the accuracy of the final results.

Ethical issues

- protection of personal data - necessary for credit score calculation
- explainability and transparency of the used ML algorithm
- introduction of bias – danger of discrimination for ethnic minorities by implicit correlation
- lack of accuracy, objectivity, and accountability of credit score computation

Solving ethical issues

- use of interpretable DM/ML algorithms/models
- preparation of training data samples to avoid bias
- protection of personal data against breaches through anonymization
- training all employees to work with DM/ML algorithms and know their biases
- continuous human supervision of DM/ML algorithms