CRISP-DM
Business Understanding
Data Understanding

PROFINIT

Petr Paščenko

2021

Outline

PROFINIT

1. CRISP-DM
2. Business Understanding
3. Data Understanding

CRISP-DM

Cross Industry Standard Process for Data Mining

PROFINIT

- Old (‘96) but there is no better
  - good shopping list for project planning
- Iterative
  - data science is a science
  - agile, before it was cool
- Data Science centric
  - not suitable for production development

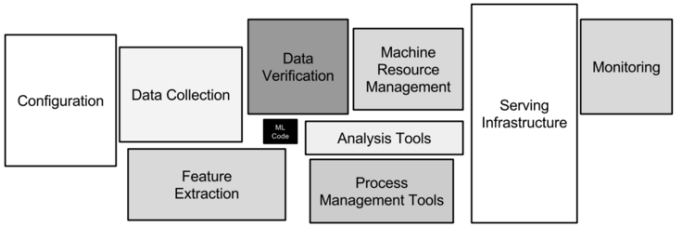
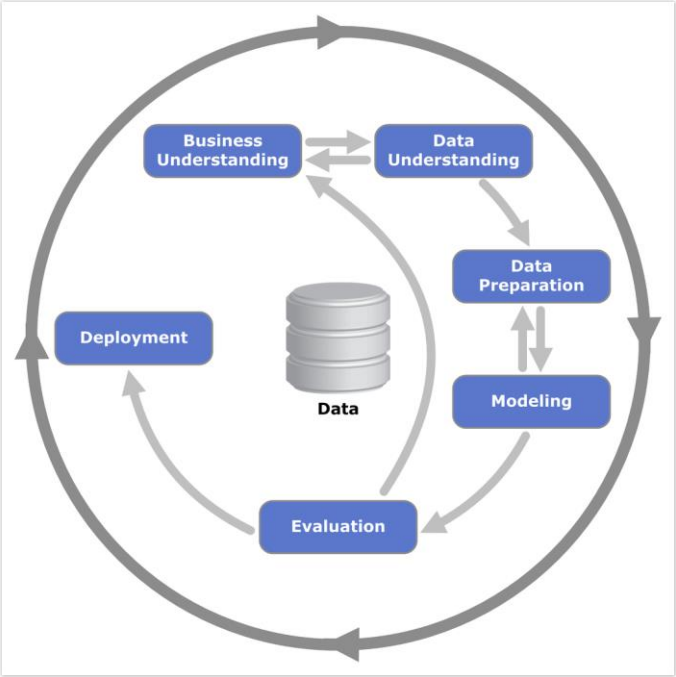


Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.



1. Business Understanding

You can tell a fool by the wine he pours next to his glass.
— Umberto Eco

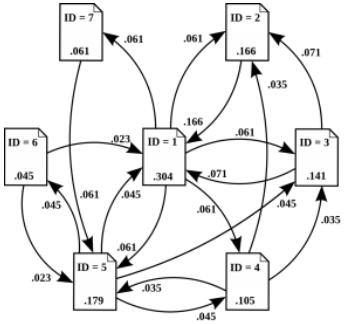
The decision

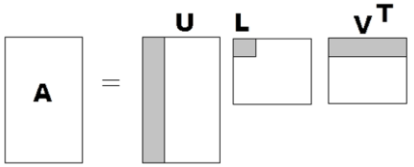
PROFINIT

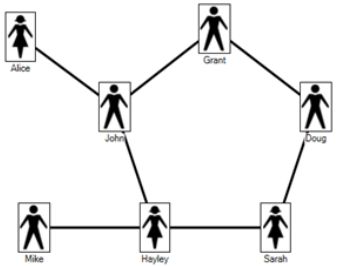
- The goal of every data science project is to enhance the decision process in a very particular point
  - Accept or reject? (student, mortgage, employee)
  - Spam or ham? (email, phone, phishing, apps)
  - Me or not me? (authentication, identity, fraud)
  - Normal or abnormal? (warning systems, smart validations)
  - Most similar other? (person, profile, document)
  - What did I mean? (search results, voice control)
  - Where am I? (spatial navigation)
- There is always a default / expert option
  - Send it to everybody
  - Discard all emails containing “Viagra”

Try yourself, what is the key decision?

PROFINIT







Key business metric

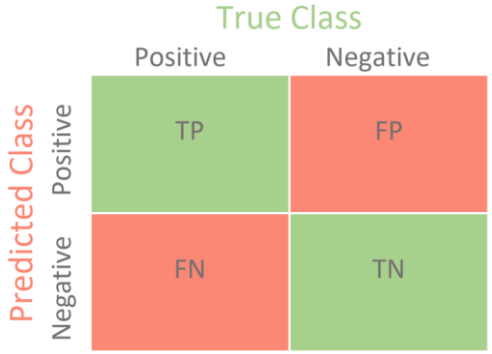
PROFINIT

- Hard number, we try to minimize to maximize
  - Click rate
  - Credit default rate
  - Time spent in the app
  - Package return ratio
  - Miles per package
  - Top 5 ratio
- Easy conversion to money
  - 1‰ decrease of default rate saves us 20M a year
- Usually differs from model error
  - RMSE, R2, accuracy, precision, AUC, log-loss, fpr, fnr,…

Key business metrics – more complex cases

PROFINIT

- TRUE/FALSE POSITIVE/NEGATIVE
  - the price of FP and FN usually differs
    - covid test: quarantine vs. possible epidemic
    - bank risk: rejected loan vs. client default
    - cancer scan: death vs. non needed chemo
- Transformation
  - Best possible TPR with FP below 1%
  - AUC, lift, gain, etc.



Non functional requirements

PROFINIT

- How quickly we must decide
  - real time decisions (autonomous driving)
  - near real time (interactive / non interactive) – e.g. online translator, navigation
  - batch systems – minutes, hours, over night, days, week, months
- Prediction horizon (forecasting vs futurology)
  - for predicting tasks – credit default, user churn, expected miles per remaining petrol
- What are the data available
  - size, quality, history, representativeness, documentation
- Interpretability
- Possible knowledge extraction
- Other limitations (security, computational resources)

Key Roles and results

PROFINIT

- Domain expert
  - understand the business domain (banking, language translation, medicine)
- Database expert
  - understand the structure and semantics of the data
- Data Analyst
  - understand what is possible to do and what is not
- Result
  - Brief memo / presentation with business summary
  - Written to be understood by all three experts
  - Covers all the mentioned questions (problem definition, data scope, business metric, functional and non functional requirements)
  - Mutual acceptance by ceartor and consumer

2. Data Understanding

Show me your data, and I'll tell you who you are.
— Native American proverb

Task

PROFINIT

- Data exploration report
  - rmd or ipynb or other
- Explore key dataset properties
  - regarding specified tasks or in general
- Key dataset properties
- Tables structure and values
- Data origin and quality
- Descriptive statistics
- Modest data visualisation

Goal:

- *exploration of dataset XY (regarding problem P)*

Data:

- *dataset XY, obtained from source Z,*
- *limited to cases ABC, from 2020 to 2021*

…

Summary:

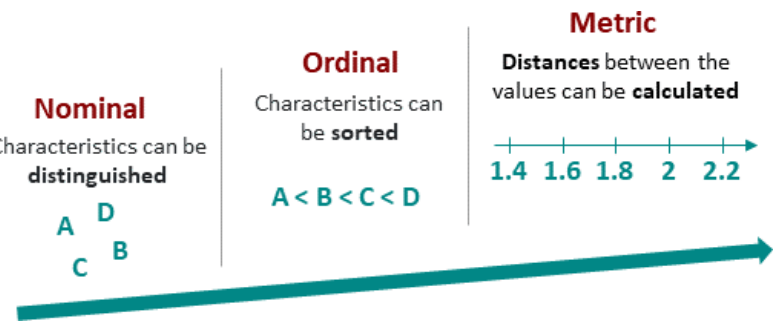
- *regarding problem P, there is no useful data in dataset XY because of reasons 1,2,3*

Key Dataset properties

PROFINIT

- Size
  - small / big data (does it fit on the RAM, HDD)
- Availability
  - who is the owner, can we access/download them, security, GDPR
- Completeness
  - which data tables covers which specific tasks, what is missing, anonymization
- Structure
  - db tables, csv, json, plaintext, encoding, binary, audiovisual, other
- Quality
  - what is the data source, original system, is it cleaned, consolidated,…
- Relevance
  - which parts are relevant to the addressed problem

Table values



- Nominal
  - domestic pet (dog, cat, other), city district
- Binary
  - M/F, indicator variables (user read book, seen movie, …), active/closed
- Ordinal
  - education (elementary < high school < college < scientific)
- Numeric
  - discrete/continuous, positive, nonnegative
- Date/Time
- Primary / foreign keys

Data origin and quality

- Exact value
  - verified, measured, estimated, other model result
- Primary vs aggregated data
  - transaction time vs transactions per day
- Missing values
  - are there missing values, how are they encoded
  - db NULL,
  - „null“, „nil“, „NA“, „none“, „N/A“, “”, “UU”, “UUU”, “Unk.”
  - 0 (integer e.g. telephone number), -1 (income), -2 …
  - 99999 (small integers i.e. children number)

Descriptive statistics

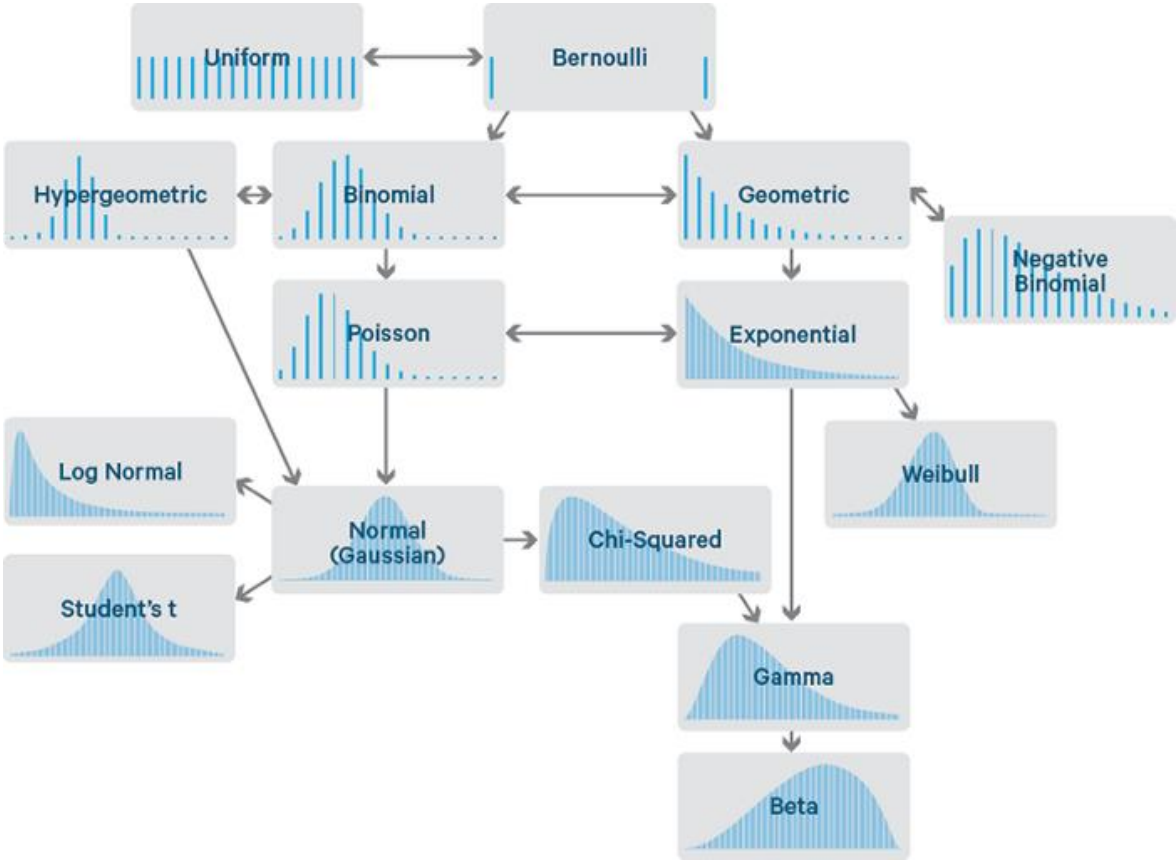
PROFINIT

- min, max
- mean / average
- median
- mode
- standard deviation
- interquartile range
- selected quantiles (0.01, 0.1, 0.9, 0.99)
- missing values ratio

Probability distribution

PROFINIT

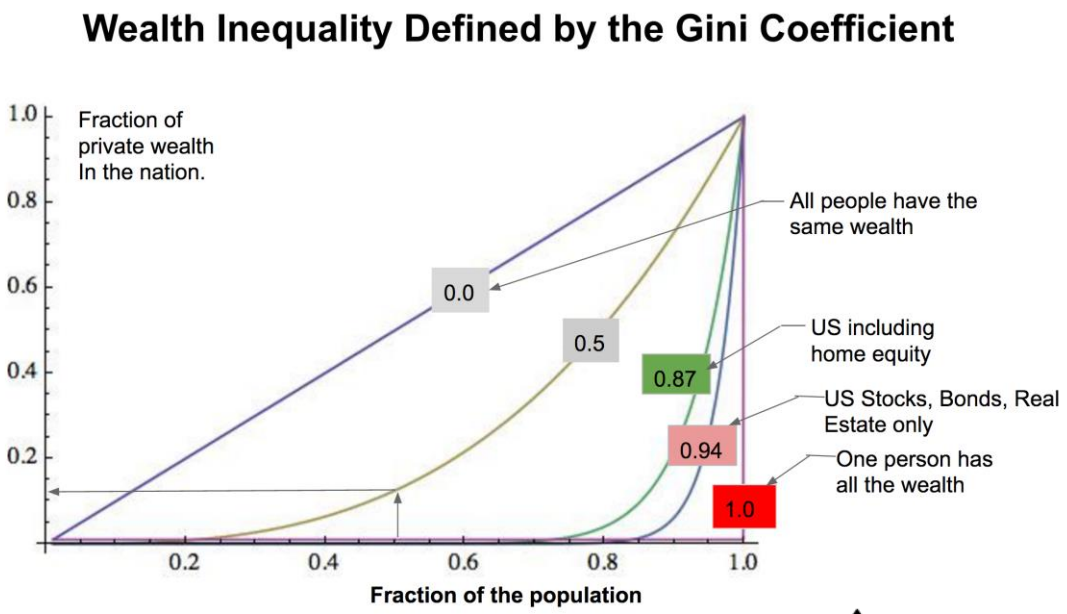
- discrete / continuous
- uniform, normal
- symmetrical
- long/heavy tail
- bimodal

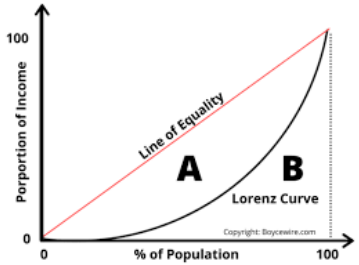


Outliers

PROFINIT

- Are there outliers?
  - very large, very small, or otherwise strange values
- Is it an error?
  - height [m]: 1.81, 1.79, 1.95, 18.7, 1.68
- Does it affect the aggregate (sum, mean)?
  - wealth [$]: Bob 120 000, Maria 250 000, Bill 250 000 000 000

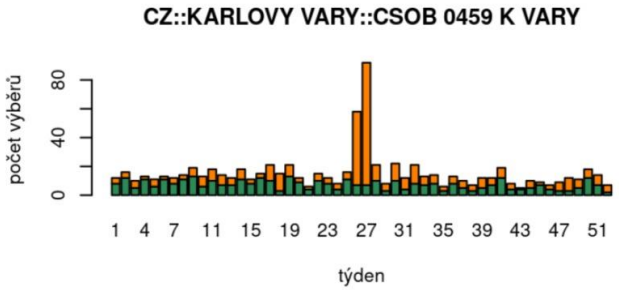


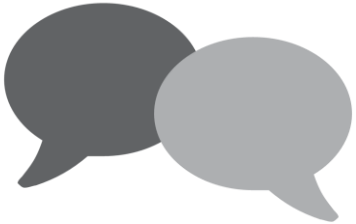


Time series

PROFINIT

- Is the series stationary
  - same distribution over time
- Is there a trend?
- Are there seasonalities or cycles?
- Are there discontinuities?
  - measurement or methodology change





Diskuze