Instructions and tips for datasets

General

Each of tasks below should be processed by CRISP-DM methodology, which will be presented on lectures. We expect an analytic report that should contain at least:

- description of the solved problem – to be obvious that you understand the task
- input data description
- process steps, methods, techniques (e. g. description of data transformation)
- data exploration with proper charts/diagrams – to be obvious that you have understood data and relations
- (possibly simple) predictive or classification model (if not stated differently) for a proper target
- model performance evaluation
- summary, conclusion

The expected volume of a report is 6 to 12 pages (recalculated to the A4 format) of text and charts. You can keep a source code in your report but it is not taken into account for the report volume.

The report may be made in Czech or in English.

Changes of instructions are possible. Student may suggest any change if there is a reason for it; teacher will deal with this demand as soon as possible.

Datasets

Datasets are saved in GitLab repository: https://gitlab.mff.cuni.cz/mlyni8am/ndbi048. Datasets have different size and complexity. At large or complex datasets the student may focus on a particular problem and work only with subset of columns and rows.

There are two datasets for demonstration purpose, not for assignment:

- titanic2.zip
- homecredit_default.zip

The list of available datasets with descriptions and tips for analysis follows.

King James Bible (bible-kjv.zip)

The Bible in an ancient English translation (17th century). Plain text where each row is one verse with structure book name, chapter:verse number, verse text.

Exploration tips:

- length of items (chapter, verse or word number) and quality (the most frequent words or word patterns including or excluding stopwords etc.) in different Bible books

Model tips:

- classification for each verse, whether it belongs to the Old Testament, or to the New Testament
- finding similar or dissimilar books

NHL Players (nhl-player-data.zip)

Individual statistics of the NHL players in seasons 2004 to 2018 and the list of teams. Description (in Czech) of columns is in the extra file README.txt.

Exploration tips:

- correlation of some player performance metric (goals, ice-time, faceoffs etc.) and other features (post, ice-time, age etc.)
- detect player's trade, explore relation to various features (age, team, goals, ice-time, previous trades etc.)
- relation of rank (as a number or as a presence among top N players) to various statistics and features in the same season

Model tips:

- prediction of performance (goals, ice-time, points etc.) in the next season either as binary (will/won't get better than before, will/won't be over a threshold) or quantitative
- prediction of player's trade in the current or the next season
- prediction whether player's rank will fit into top N

Chess ratings (chess_ratings_upravene.zip)

Official monthly lists of chess players in standard (Dec 2017 to Dec 2022), rapid and blitz games (Jan 2022 to Dec 2022). Each file contains the same set of columns: Player ID, name, federation (country), sex, general title (if players holds it, e. g. GM = grandmaster), woman title, other title (e. g. trainer or arbiter), title in online games, current rating, count of played games in previous month, coefficient of development (unit of rating change – high for newcomers and low for top players), birth year, flag (e. g. "i" = inactive), current year and current month.

Exploration tips:

- distribution of rating for players with various titles
- changes of counts of players over rating R
- rating and played games for different countries, increase and decrease tendencies; is there generally inflation or deflation of the rating?
- distribution of monthly and yearly changes
- relationship of games played and month
- how do covid years (2021, 2022) differ from others; when did it return to normal?
- frequency of online and other titles, relationship to rating and to general titles

Model tips:

- prediction of player rating next month or a year after (in absolute measure or binary "will be higher or not")
- prediction of number of games played by individual player
- prediction of average rating height/change for top N players overall and in some special countries (India, China, Vietnam, USA, UK etc.)

Card transactions at petrol stations (card_petrol.zip)

Card transactions made by clients of some small Czech bank at petrol stations, mostly in the Czech Republic but also abroad, in the period of 12 months in 2017-2018. Description of columns is in the extra file README.txt.

Exploration tips:

- distribution of clients by number of payments, number of distinct stations, average amount etc.
- distribution of amounts and relationship to season, hour, day in month etc.
- distribution of stations by number of payments, number of distinct clients, average amount etc.
- distribution of distance client-station, relationship to season, hour, day etc.
- distribution of payments during a day, a week, a month
- trends (time series) of payment number, amount etc.
- distribution of stations in the Czech Republic, highway twins (stations on opposite sides of highways) similarity/dissimilarity
- favourite station brands

Model tips:

- prediction of payment number or total amount for a station (daily, weekly, during night etc.)
- prediction of payment number or total amount for a client (monthly)

Jokes (jokes.zip)

Evaluation of 100 English jokes by many readers. The zip file contains two Excel tables with evaluations and a directory with jokes texts in separate HTML files. In each Excel table, rows are readers, the first column is a number of evaluated jokes by the reader and other 100 columns contain evaluation from −10 to 10 points (possibly decimal, 99 = unevaluated) for jokes 1 to 100.

Exploration tips:

- distribution of joke evaluation (individual and as average over readers) by jokes and by readers
- features of jokes (length, frequent words etc.) and correlation to the evaluation
- correlation of evaluation by readers (reader A likes/dislikes similar jokes as reader B) and by jokes (joke X is liked by similar readers as joke Y)
- jokes with high and low variances in evaluations

Model tips:

- prediction of average joke evaluation, of variance or of readers count (when the joke is quite new for all)
- prediction of individual evaluation of particular joke by particular reader

Tennis matches (tennis_matches.csv)

Statistics of sampled tennis matches (both men and women) in the period 2015–2023. Description of columns is in the file matches_data_dictionary.txt.

Exploration tips:

- share of won games for players by surface – are all players universal, or is the surface important for some of them?
- individual metrics of play quality (i. e. share of doublefaults) – are some players special?
- correlation of match result and ranking before the match, with possible relationship to surface
- “favourite” opponents – highly unbalanced long-term score between rather equally ranked players

Model tips:

- prediction of the next match result
- prediction of match feature value (duration, number of games, number of aces etc.)

Brazil accidents (brazil_nehody.csv)

Data on traffic accidents in Brazil from 2007 to 2023 (end of June). The zip file contains one csv file for each year, where each row is one accident and columns are descriptors of the accident (e. g. day, time, weather condition, GPS or number of injured). Column names are self-explanatory but in Portuguese, as well as texts in the dataset. Additionally a file with road radar places is included.

Exploration tips:

- number of accidents by day/hour of week, by season (month or special periods like Christmas)
- trends in number of accidents overall and for special places/towns/roads
- correlation of accident indicators like number of dead and injured, hour, day of week, weather condition etc.

Model tips:

- prediction of number of accident for main roads/cities and next day (N days, month etc.) – with or without knowledge of weather condition
- detection of new accident places or other breakpoints in trends

Elections results in the Czech Republic (elections.zip)

Results of parliamentary elections in 2013 and 2017 years by election districts (the smallest unit for which votes are summed) together with some sociodemographic descriptors of the district population (e. g. share of men/women, share of unemployed people, share of children).

The zip file contains two data files with identical content (one in csv format and one in SPSS format) and two files with description of data and columns. For some explorations, it may be useful to exploit GPS coordinates of the municipalities – see Population of Czech municipalities dataset.

Exploration tips:

- distribution of sociodemographic indicators in districts, municipalities, counties, differences by size of the municipality
- relationship of vote share by individual parties and sociodemographic indicators
- distribution of vote share changes from 2013 to 2017 in districts, municipalities, relationship to sociodemographic indicators
- distribution of turnout (in %)
- anomalities, places different from their surrounding

Model tips:

- prediction of vote share for party X in district/municipality Y in 2017
- prediction of turnout (in absolute numbers or binary “over/below total average”

Population changes in all world’s countries (populace_svet.zip)

Population and other demographic data on every country (so called “economic”) in the world from 2000 to 2021. The zip file contains individual csv files for various demographic indicators (rows=countries, columns=years), an overview table and country groupings with

respect to different categories. Some indicators are obvious, some need to be discovered or understood and compared to external sources.

Exploration tips:

- population, median age or life expectancy trends in groups
- correlation of population indicators
- which countries (kind of them) had growing and which had decreasing population or other indicators
- typical pattern of growth/decreasing/stagnation overall and in region, atypical cases for region
- changes in fertility, mortality etc. during 2000–2021 period overall, by groups or by countries

Model tips:

- prediction of some population indicator for next N years by country or by group (in absolute numbers or binary "will/won't be higher")

Population in Czech municipalities (populace_obce_cr.zip)

Population and its changes in all Czech municipalities yearly from 1971 to 2020. The zip file contains 77 excel files, one for Prague and others for individual (former) districts. Each excel file contains information on population state and changes for each municipality and each year from 1971 to 2020 (if the municipality itself existed in that year). Additionally, a csv file with GPS coordinates for municipalities is added (note: the encoding of this csv differs from excel files).

Exploration tips:

- population trends in the whole country and by regions/districts
- which municipalities (kind of them) were growing and which were decreasing
- typical pattern of growth/decreasing/stagnation, atypical cases
- municipalities with growth/decreasing due to migration vs. due to newborns/deaths
- migration from villages to towns and back

Model tips:

- prediction of population N years ahead for individual municipalities, districts (in absolute numbers or binary "will/won't be higher")
- prediction of newborns and deaths for next year

Song lyrics (lyrics.zip)

List of popular songs with lyrics. Some rows are music composition without lyrics but we treat all dataset as "songs". The dataset contains five columns without header: id, song name, year of publishing, interpret name and assigned genre. There are hyphens instead of spaces in names of songs and interprets.

Exploration tips:

- genres and interprets with most songs
- multigenre interprets
- distribution of songs features (word count, number of unique words etc.)
- song covers and remakes (same name and lyrics, possibly with small differences)
- typical words or patterns for various genres and for various interprets

Model tips:

- classification of song to genre according to lyrics
- classification of song to interpret according to lyrics

Stopwords (stopwords.zip)

English “stopwords”, i. e. frequent words without a special meaning, like “and”, “you”, “how” etc. For auxiliary use in text processing.

Some datasets may be linked with other public data. We expect that student can do it individually.