MUNI
ECON

Artificial Intelligence in Finance

Introduction - part B

Štefan Lyócsa

Department of Finance, Faculty of Economics and Administration

October 3, 2024

Outline for Section 1

Data structures

Missing observations

Outliers

Data transformation

Feature engineering - part A

Other data considerations

Datasets

Data structures

How much **effort** (time) used, in a data-related project, consumes data collection and data pre-processing?

Data structures

How much **effort** (time) from the start to the finish of the problem solving, is taken by data collection?

from 50% to 80% [6].

Guess (2016, [2]) reports results from a survey from CrowdFlower:

- 60% cleaning and organizing data,
- 19% collecting data,
- 9% modelling & machine learning,
- 4% refining algorithms,
- 3% building training data sets,
- 5% other.

Data structures

Depending on the problem at hand, we work with four data structures:

1. **cross sectional** independent units - all observations are assumed to be retrieved at the same time/moment.
2. cross sectional **dependent** (spatial) units - observations are dependent (etc., geographically, households, families, ...).
3. **time-series** units - observations are ordered according to time.
4. combination of previous structures.

Some other notable data (obs. unit) type challenges:

- **Multiple dependent** variables?
- Unobserved, **latent**, variables of interest?

Outline for Section 2

Data structures

Missing observations

Outliers

Data transformation

Feature engineering - part A

Other data considerations

Datasets

Missing observations

Data collection often leads to **missing observations**:

- We can miss **full units** of observations → **sample selection bias** (see Heckman [5]), e.g.:
  - Non-response bias in surveys.
  - Application criteria.
- We only **miss certain attribute** of a unit, e.g. we do not observe the age or the gender of a customer.

The later is of interest today.

Missing observations

Here we have a snapshot of a dataset with characteristics of apartments in Prague:

cena	cenam2	m2	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	d1kk	d11	d2kk	d21	d3kk	d31	d4k	cihla	novostava	porekon	dobry
4.00	142.86	28	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1
4.30	153.57	28	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0
4.36	155.68	28	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0		1	0	0
4.41	157.50	28	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0			0	0
4.49	104.42	43	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0
4.50	155.17	29	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1	1		0
4.50	150.01	30	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1		0
4.50	128.57	35	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0
4.59	114.75	40	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	

Missing observations are highlighted with red.

- Should we even care?
- What to do with missing observations?

The **choice** of a **specific approach** to missing data depends on the **reasons behind** the missing values.

Missing observations

Assume that the parameter of interest is θ (e.g. credit score, profit). Missing data can be classified as [10]:

- **Missing completely at random** (MCAR) suggests, that there are no systematic differences between missing values. Alternatively, the estimate of θ is independent of whether data are missing or not.
- **Missing at random** (MAR) suggests, that part of the missingness can be explained by **known** variables. Alternatively, missingness is conditionally independent of the estimate θ.
- **Missing not at random** (MNAR) suggest that part of the missingness can be explained by **unknown** or **not measured** variables.

Missing observations

List-wise deletion

If we assume missing completely at random (MCAR), we can remove all units that have a missing value, perform **list-wise deletion**. However, in some instances, this can lead to drastic reductions, e.g.:

cena	cenam2	m2	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	d1kk	d11	d2kk	d21	d3kk	d31	d4k	cihla	novostava	porekon	dobry
4.00	142.86	28	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1
4.30	153.57	28	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0
~~4.36~~	~~155.68~~	~~28~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~1~~	~~0~~	~~0~~	~~1~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~		~~1~~	~~0~~	~~0~~
~~4.41~~	~~157.50~~	~~28~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~1~~	~~0~~	~~1~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~			~~0~~	~~0~~
4.49	104.42	43	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0
~~4.50~~	~~155.17~~	~~29~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~1~~	~~0~~	~~1~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~1~~	~~1~~		~~0~~
~~4.50~~	~~150.01~~	~~30~~	~~0~~	~~0~~	~~0~~	~~0~~	~~1~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~1~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~1~~	~~1~~		~~0~~
4.50	128.57	35	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0
~~4.59~~	~~114.75~~	~~40~~	~~0~~	~~0~~	~~0~~	~~0~~	~~1~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~1~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~1~~	~~1~~	~~0~~	

Only 2.77% (6) of data-points are missing (out of 216), but we remove 5 units (rows) or 52.7% (114) of all data-points. Huge **sacrifice** (not a good trade-off) if you ask me.

Missing observations

Single imputation methods

You impute a single value, e.g. [8]:

1. **Random** imputation ignores potential patterns in missingness and imputes a random value from a possible range of values or from a given probability distribution.
2. **Mean/median** imputation substitutes the unconditional mean (continuous variable) or median (for dummy variables).
3. **Match-based** imputation.
   - **hot-deck** imputation substitutes the missing value with one from a similar unit from the **same** dataset.
   - **cold-deck** imputation substitutes the missing value with one from a similar unit from a **different** dataset.
4. **Predictive** (model based) imputations is based on a statistical model (regression, random forest,...). To be discussed later.

Missing observations

Multiple imputation methods

Single imputation methods **do not assume errors** in the predictions of the missing values. An alternative is to create **multiple datasets**. A possible procedure is as follows [10]:

1. Start with an initial dataset $Z^{b=1}$ with missing values and $k = 1, 2, ..., p$ features.
2. Perform single imputation (random, mean/median).
3. For each feature $k = 1, 2, ..., p$:
   - Estimate an imputation model M_k.
   - Use model M_k to predict the value of the missing observations of the k^{th} feature.
4. Save the new dataset $Z^{*,b=1}$ with no missing values.
5. Use appropriate **re-sampling** method to $Z^{b=1}$ and repeat steps 2 and 4 until you have $b = 1, 2, ..., B$ datasets $(Z^{*,1}, Z^{*,2}, ..., Z^{*,B})$.

Outline for Section 3

Data structures

Missing observations

Outliers

Data transformation

Feature engineering - part A

Other data considerations

Datasets

Outliers

The issue:

- A detailed definition of an outlier requires quite specific assumptions about the underling data (e.g. distributional assumptions).
- A more **general** approach views outliers as data point(s) that is (are) significantly different from other observations within a dataset [8].
- Outliers might be **valid** data (from a different distribution), but also **mistakes**, which makes identification complicated.

Outlier

Grubbs's

- If data are from **normal distribution** (a dream you should rarely assume) you can use **Grubbs' test** [3]. Let $X_i, i = 1, 2, ..., n$ denote observations from a normal distribution. The H_0 (null hypothesis) of no outlier is tested as:

$$ESD = \max_{i=1,2,...n} \frac{|X_i - \bar{X}|}{s} \tag{1}$$

with s being the sample standard deviation. The critical value is given via Student's t-distribution.

Outlier

Hampel identifier/filter

- **Non-parametric** approach to label '*potential*' outliers:

$$R_i = |X_i - \tilde{X}| \tag{2}$$

$$MAD = \tilde{R} \tag{3}$$

An unbiased estimate of the standard deviation for Gaussian data is found after scaling [9]:

$$MADN = \frac{MAD}{0.6745} \tag{4}$$

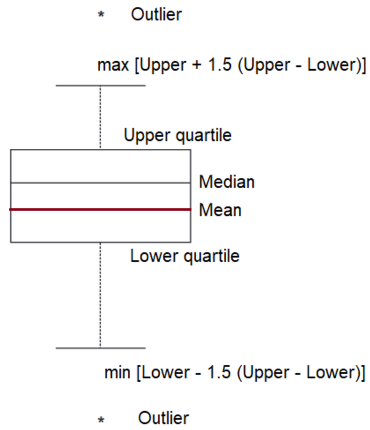
Given significance level α, **a potential outlier** X_i **meets** the following:

$$H_i = \frac{R_i}{MADN} > \sqrt{\chi^2_{1-\alpha/2,1}} \tag{5}$$

Outlier

Box-plot rule

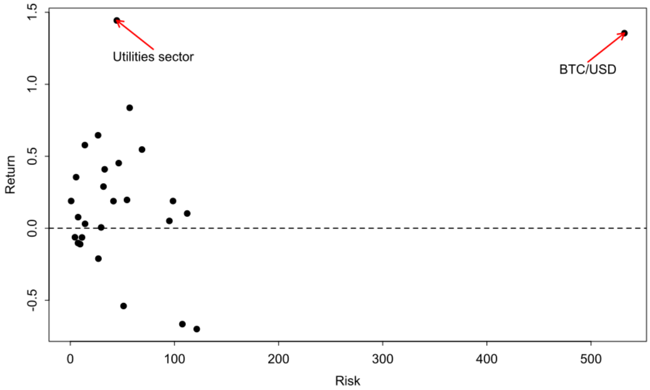
- A popular method (rule of thumb) to identify outliers is to use the **box-plot rule** (e.g. [7]):



Outlier

Multivariate outliers

- Asset from utilities sector is a likely 'return' outlier.
- BTC/USD is likely an outlier from both return and risk perspective.



Outlier

Further issues

- Inspect each continuous variable if possible.
- Be aware of the **masking effect**, which happens when there is a group of outliers; as the outlier is not alone, they mask each other.
- Alongside of testing, consider:
  - **Trimming** - removing observations, i.e. everything above the 99.99% percentile is removed.
  - **Winsorization** - substituing extremes, i.e. everything above the 99.99% percentiles is substituted with the 99.99% percentile.
  - data transformation (next section).
- If data are susceptible to outliers (market risk measures), use methods that are less affected by the presence of outliers.

Outline for Section 4

Data structures

Missing observations

Outliers

Data transformation

Feature engineering - part A

Other data considerations

Datasets

Data transformation

Conversion to dummy variable

Let $X_i, i = 1, 2, ..., n$ denote the size of the apartment in m^2. You would like to understand price setting Y_i, represented by rent (CZK) per m^2. Standard linear regression yields estimates:

$$\hat{Y}_i = 581.13 - 7.2\hat{X}_i \tag{6}$$

We could **introduce non-linearity** by converting X_i into dummies. This type of conversion is simple and variables are easy to interpret. However, such transformation might lead to an excessive increase in the number of variables.

Data transformation

Conversion to dummy variable

Let $Q(X,k)$ be returning k^{th} quintile and $I(.)$ be a signalling function returning 1 if the condition holds and 0 otherwise:

$$
\begin{aligned}
X_{1,i} &= I(X_i \leq Q(X,1)) \\
X_{2,i} &= I(X_i > Q(X,1) \wedge X_i \leq Q(X,2)) \\
X_{3,i} &= I(X_i > Q(X,2) \wedge X_i \leq Q(X,3)) \\
X_{4,i} &= I(X_i > Q(X,3) \wedge X_i \leq Q(X,4)) \\
X_{5,i} &= I(X_i > Q(X,4))
\end{aligned} \tag{7}
$$

The estimates from a linear model are:

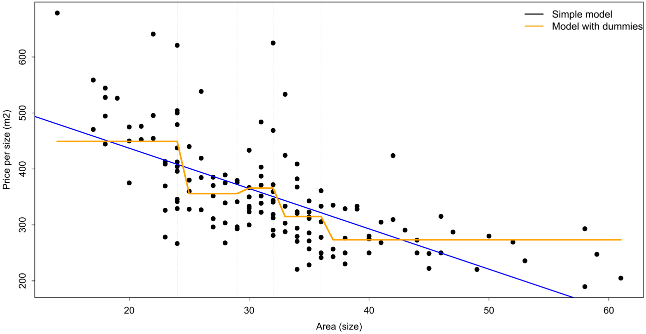
$$\hat{Y}_i = 273.47 + 175.74\hat{X}_{1,i} + 82.49\hat{X}_{2,i} + 92.19\hat{X}_{3,i} + 41.44\hat{X}_{4,i} \tag{8}$$

Data transformation

Conversion to dummy variable

$\hat{Y}_i = 273.47 + 175.74\hat{X}_{1,i} + 82.49\hat{X}_{2,i} + 92.19\hat{X}_{3,i} + 41.44\hat{X}_{4,i}$

The two models can be visualized:



Data transformation

Binning, data bucketing

Similar to the approach before is **binning**, where instead of a $1/0$ dummy a representative value is used. Continuing the example before, for the first 'bin', the values would be:

$$X_i = \begin{cases} \left[\sum_{i=1}^{n} I(X_i \leq Q(X,1))\right]^{-1} \sum_{i=1}^{n} X_i \times I(X \leq Q(X,1)) & X_i \leq Q(X,1) \\ 0 & X_i > Q(X,1) \end{cases}$$

In this case, binning and using dummies leads to the same model.

Data transformation

Smoothing

Noise in data may refer to random fluctuations around the **signal**. Some applications:

- Asset prices (bid-ask spread, liquidity, lot size constraints, decimal places,...).
- Measurement uncertainty (google trends data, surveys,...).

The idea of smoothing is to mitigate the effect of noise and recover the signal. **Methods for time-series**:

- Rolling median & mean.
- Kálmán filter (more advanced - will not cover here).
- Extracting deterministic trends, see [1, 4].

Data transformation

Smoothing: Rolling mean

Let $X_t, t = K, K+1, ..., T$ denote a time-series and $K \in \mathbb{N}$ being the smoothing window size parameter. The rolling mean:

$$Y_t(K) = K^{-1} \sum_{j=t-K+1}^{t} X_j \tag{9}$$

Data transformation

Smoothing: Rolling mean with exponential weights

Let $\delta \in [0,1]$ be a **memory parameter**, and the vector of weights is given as:

$$w_q(K,\delta) = \delta^q \left[\sum_{r=1}^{K} \delta^r\right]^{-1}$$

exponential smoothing can be expressed as:

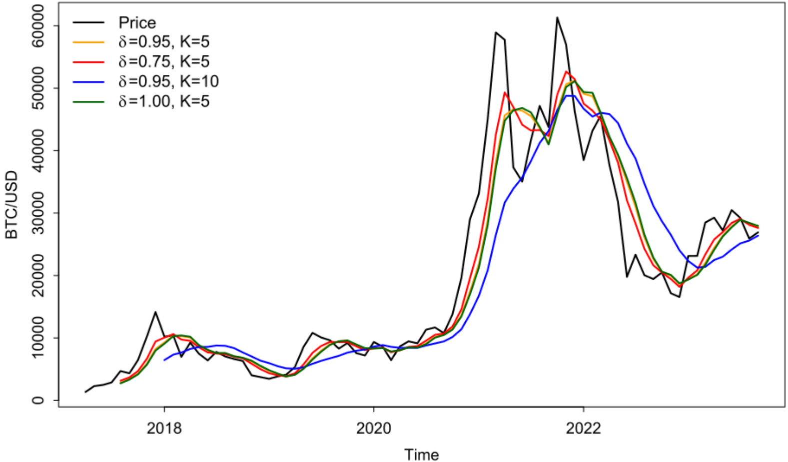
$$Y_t(K,\delta) = \sum_{j=t-K+1}^{t} X_j w_{t-j+1}(K,\delta) \tag{10}$$

What happens if we let $K = 1$? What happens if we $\uparrow K$? What happens if $\delta \to 1$?

Data transformation

Smoothing: Rolling mean with exponential weights

Let's take a look:



Data transformation

Data standardization

Data standardization is performed to make variables similar in scale or to achieve some desired data property:

- Decimal scaling.
- Z-score.
- Min-Max normalization.
- Box-Cox transformation.

Data transformation

Decimal scaling

Let $X_i, i = 1, 2, ..., n$ be the original variable and $c \in R$ a constant. Scaled variable derived by **decimal scaling** is achieved by multiplying each value using the scaling constant 10^j, where j satisfies [8]:

$$X_i^{(s)} = 10^j \times X_i,\ max_i|X_i^{(s)}| \leq c \tag{11}$$

Data transformation

Z-score normalization

Let $X_i, i = 1, 2, ..., n$ be the raw variable, $\bar{X}$ the average and σ_{X_i} standard deviation. **Z-score** scaling is achieved by:

$$X_i^{(s)} = \frac{X_i - \bar{X}}{\sigma_{X_i}} \tag{12}$$

The $\bar{X}_i^{(s)} = 0$ and $\sigma_{X_i^{(s)}} = 1$.

- Popular standardization.
- It **might change time-series properties**. (cond. heteroscedasticity changes).

Data transformation

Min-Max normalization

Let $X_i, i = 1, 2, ..., n$ be the raw variable, min_{X_i} and max_{X_i} the corresponding minimum and maximum values, and U and L the new maximum and minimum. The **Min-Max transformation** leads to [8]:

$$X_i^{(s)} = \frac{X_i - min_{X_i}}{max_{X_i} - min_{X_i}} \times (U - L) + L \tag{13}$$

It might distort the time-series properties.

Data transformation

Box-Cox transformation

Let $X_i, i = 1, 2, ..., n$, be the raw variable and λ a transformation parameter (with $\lambda = 1$ essentially untransformed variable).

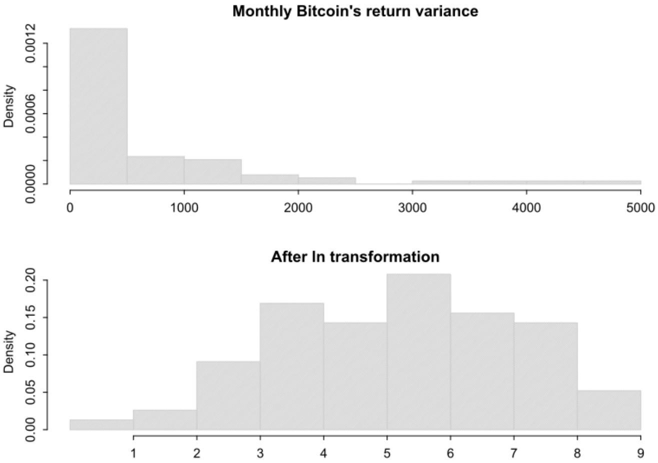
$$X_i^{(s)} = \begin{cases} \frac{X_i^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ ln(X_i) & \lambda = 0 \end{cases}$$

- Transformations can mitigate the size of extreme observations - asymmetric distributions, common in the literature.
- Sometimes $ln(X_i + 1)$ is used.
- It **might change time-series properties**.
- I use the *ln* transformations for right-skewed distributions **a lot**.

Data transformation

Box-Cox transformation

Let's compare the distributions:



Outline for Section 5

Data structures

Missing observations

Outliers

Data transformation

Feature engineering - part A

Other data considerations

Datasets

Feature engineering

Feature engineering involves three major decisions:

1. Feature **selection** - what variables to chose?
   - Curse of dimensionality.
2. Feature **extraction** - how to combine variables?
3. Feature **creation** - involves lot of creativity.
   - averages in time-series,
   - calendar effects,
   - adding ratios,
   - creating dummies (non-linear transformation),
   - de-trending, etc.

Outline for Section 6

Data structures

Missing observations

Outliers

Data transformation

Feature engineering - part A

Other data considerations

Datasets

Other data considerations

- Naming conventions.
- Exclude variable that highly correlate with others (bi-variate correlations).
  - Pearson's correlation.
  - Spearman's correlation.
  - Kendall τ correlations.
- How much data should we have? Hardware constraints and power of the tests.
- Ethical consideration when working with data.

Outline for Section 7

Data structures

Missing observations

Outliers

Data transformation

Feature engineering - part A

Other data considerations

Datasets

Datasets

- Cross-sectional:
  - Offered rental price on apartments.
  - Offered price for apartments.
  - Price of used cars: different models.
  - Credit risk on loans.
  - Household income and expenses.
  - Profitability of a business.
- Time-series:
  - Unemployment rate, GDP growth.
  - Oil and Gold price.
  - Stock price variation.

[1] Jushan Bai and Pierre Perron. “Computation and analysis of multiple structural change models”. In: *Journal of applied econometrics* 18.1 (2003), pp. 1–22.

[2] *Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says*. https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/. Accessed: 2023-10-02.

[3] Frank E Grubbs. “Sample criteria for testing outlying observations”. In: *The Annals of Mathematical Statistics* (1950), pp. 27–58.

[4] Alastair R Hall, Denise R Osborn, and Nikolaos Sakkas. “Inference on structural breaks using information criteria”. In: *The Manchester School* 81 (2013), pp. 54–81.

[5] James J Heckman. “Sample selection bias as a specification error”. In: *Econometrica: Journal of the econometric society* (1979), pp. 153–161.

[6] R Karthik and S Abhishek. “Machine Learning Using R: With Time Series and Industry-Based Use Cases in R”. In: *Apress* 2.321 (2019), p. 1.

[7] Sang Kyu Kwak and Jong Hae Kim. “Statistical data preparation: management of missing values and outliers”. In: *Korean journal of anesthesiology* 70.4 (2017), pp. 407–411.

[8] Fred Nwanganga and Mike Chapple. *Practical machine learning in R*. John Wiley & Sons, 2020.

[9] Ronald K Pearson et al. “Generalized hampel filters”. In: *EURASIP Journal on Advances in Signal Processing* 2016 (2016), pp. 1–18.

[10] Matt Wiley and Joshua F Wiley. *Advanced R Statistical Programming and Data Models*. Springer, 2019.

MUNI
ECON

Artificial Intelligence in Finance

Introduction - part B

Štefan Lyócsa

Department of Finance, Faculty of Economics and Administration

October 3, 2024

MASARYK
UNIVERSITY