Data & Science

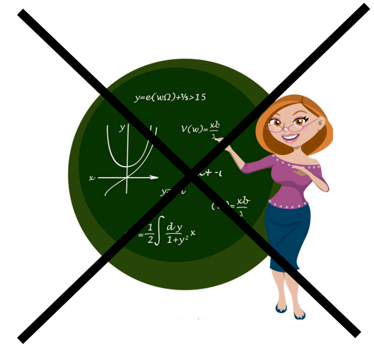
A mandate for data driven corporate innovation

By Igor Stojković
Enterprise Analytics & Data
Phillip Morris International

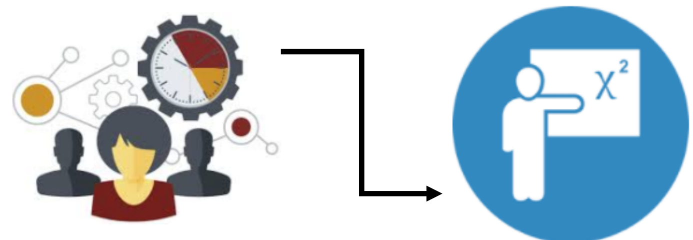
Contents

- Mathematics for data science in commercial environment
  - To prove or not to prove
- Multidisciplinary teams and Agile
- Rlabs at ABNAMRO
- Transforming discussions with business stakeholders into mathematical models
  - Business & Data understanding/experiment design/data prep/modeling/performance valuation
- Second hand car sales model
  - Kalman filter
  - Long term short term memory (LSTM) neural network model

Mathematics for data science in corporate environments



- Not about proving rigorous statements (☹)
  - Deductive vs inductive science
- Willingness to dive into business details and mathematicise them



- Creative analytical thought
  - Apply advanced techniques in novel ways for operational excellence, new markets and products
- Keep reading papers all the time
  - My current reading: Wasserstein Generative Adversarial Networks (WGAN)
  - Don’t get bored because it will kill you!

Multidisciplinary teams-Agile

Senior Stakeholders

- Accept or reject proposals

Product owner

- Determines what needs be built

Scrum Master

- Guards the process

Development Team

Data Scientist

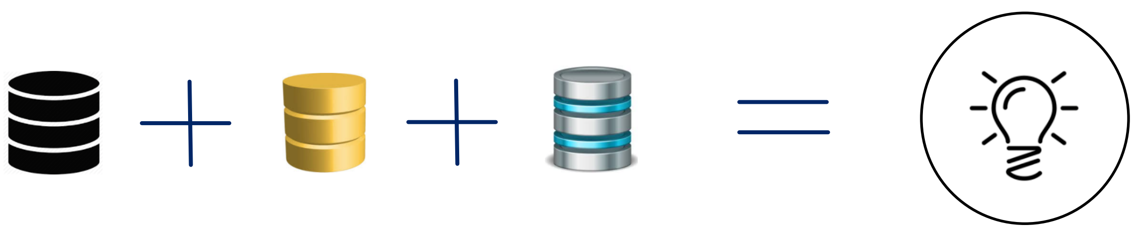
Domain Expert

Data Engineer/Hunter

A Data Science objective: Rlabs@ABNAMRO Bank

- Risk as a Service (RaaS)
  - Combine internal credit risk management knowledge with data&science to build new API services for internal and external usage
    - More efficient and up to date risk management
    - New proposition to clients

- Utilize internal and external data sources



- Consider different sub-sectors separately

How to approach??

- A general observation:
  - A washing service SME serving hotels is not interested in PD, LGD, EAD (Basel) CR models
  - Is interested in predictions on number of sold beds per hotel
    - Steering their business
  - Such models are a novelty in banking industry and valuable for risk management

- Collected domain expertize and requirements through internal and external discussions:
  - Which operational figures are crucial about performance of an SME active (e.g. a hotel), that is relevant to creditors as well as buyers and/or suppliers of entities considered?

- Boundaries
  - External information availability/price of data sources
  - Privacy

Dutch second hand car dealership forecast model

- Goal: sales forecasts at postal code area level (4 digits)

- Available sales events with
  - Car specs
  - Car age
  - Quantity sold
  - Dealer’s & consumer’s postal code

- Other available data:
  - Martkplaats data with average prices per car specs/age/period
  - Internal data on consumer behavior (aggregated to areas’ level)
  - APK data

First modeling steps

- Data prep
  - Cleaning – sounds trivial but can be extremely time consuming or even require deep modeling itself
  - Transforming data structure: aggregate, merge, find suitable representations – sometimes deeply analytical

- Target design
  - # cars sold per period, postal code area, price class & car age
  - Price classes determined by clustering

- Model design choices
  - Kalman filter
  - LSTM model

Predictive features design

- PC area of dealer and consumer
  - Where do clients of car dealers live (distribution)
- Consumer behavior contains clues about driving patterns at PC level
  - Second hand and new car ownership incidence
- APK data contains information on car decay incidence
  - How often do owners change their second hand cars

Klaman filter solution details

$Y_t := X_t * \alpha_t + \varepsilon_t^Y, \quad \varepsilon_t^Y \sim N(0, \Sigma_Y), \quad X_t - predictive\ features\ known\ at\ t$

$\alpha_t := F * \alpha_{t-1} + \varepsilon_t^\alpha, \quad \varepsilon_t^\alpha \sim N(0, \Sigma_\alpha),$

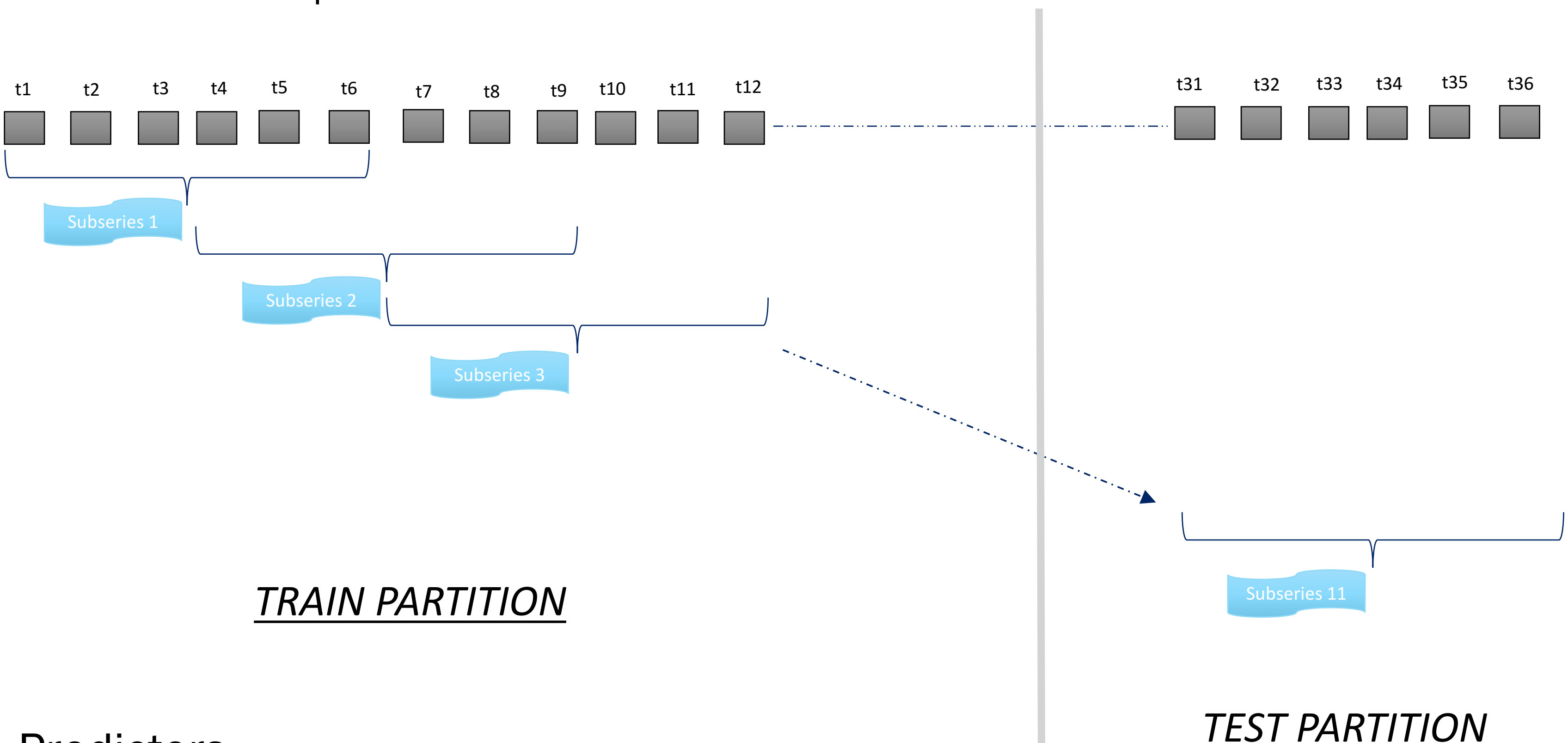
Σ_Y, Σ_α - unknown covariance matrices
F- unknown matrix to be estimated
This is a generalization of the local level model.

- 3000 time series each with a 6 month horizon
- Neighboring observations have a 3 months overlap
- In total 36 time points per time series

- Application of embedding layer technique significantly enhanced performance
  - We clustered PC's vector representations and trained Kalman filter parameters per cluster (iteratively, passing results at end of an epoch as input to the next epoch within a cluster)

LSTM neural network

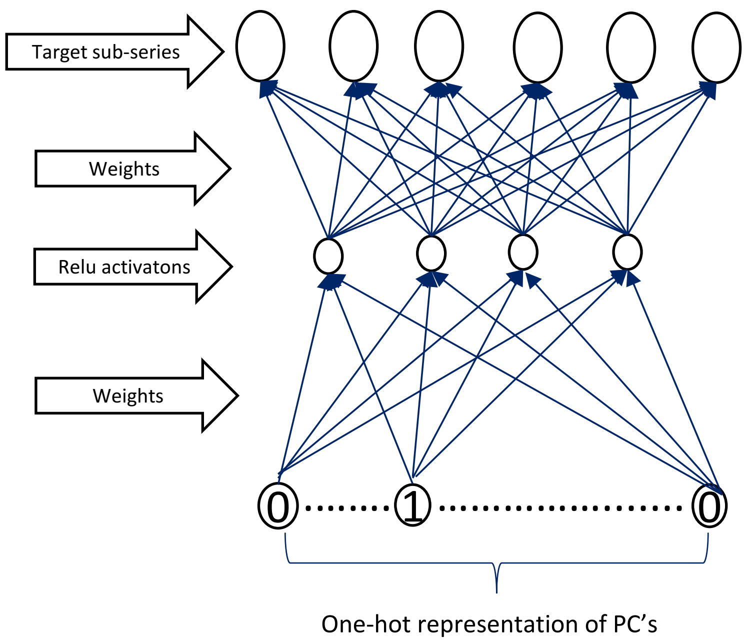
- Target redesign
  - ‘Cut up’ 36 points series (6-8 points per new observation)
  - Gives multiple observations per series
  - Some overlap is ok but not too much

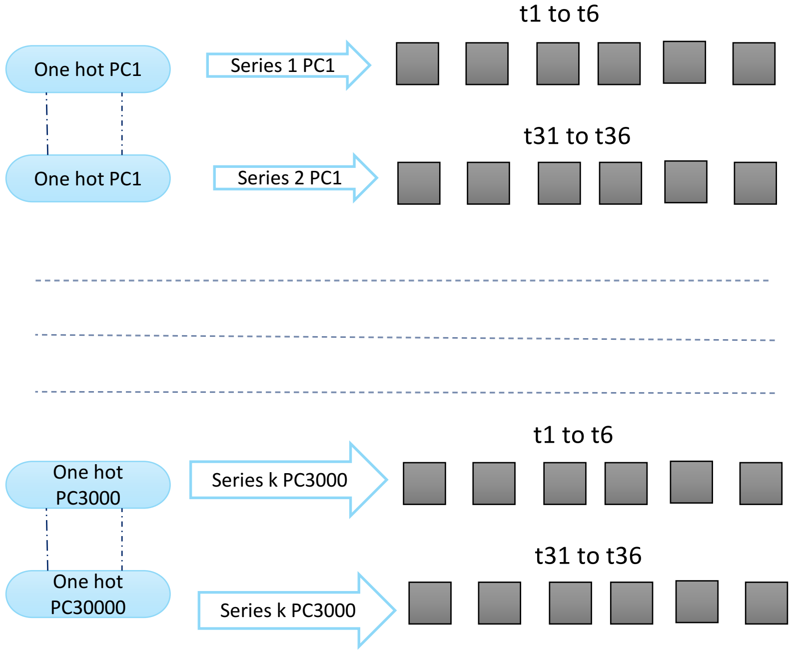


- Predictors
  - Original features series plus embedding layer values

Embedding layer

- We train a simple NN with one hot's of PC's as inputs and series parts (c.q. 6 quarters) as target values
- Hidden layer gives a vector representation of abstract PC ids in relation with its series behavior





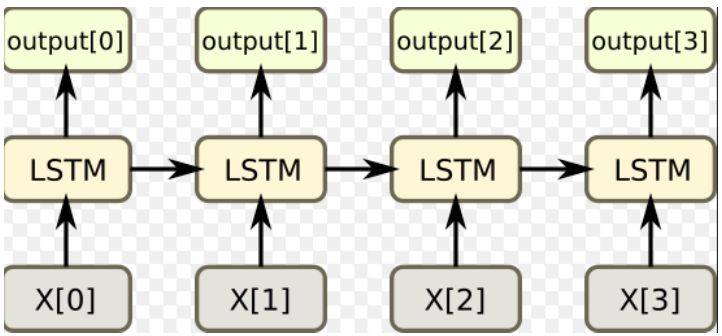
Embedding layer model formulation

- $h(x) := \sigma(W_1 * \mathrm{x} + w_1)$, x – one hot representation of a PC area, W_1 and w_1 weights of the hidden layer
- t(h):=$\sigma(W_2 * \mathrm{h} + w_2)$, W_2 and w_2 are weights of the output layer
- $\sigma(z) \coloneqq (z_1^+, \ldots, z_k^+), for\, z \in \mathbb{R}^k$
- (W_1, w_1, W_2, w_2):= $\mathrm{E}(s - t(h(x)))^2,\ \ s\ is\ target\ series,$ $E\ is\ taken\ w.r.t.data$
- Features to add to LSTM model or to use for clustering series for joint Kalman filter inference:

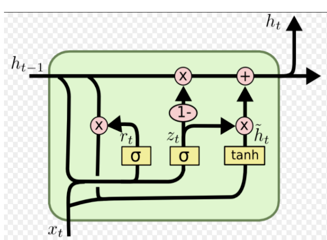
$$W_1 * \mathrm{x} + w_1 \quad (\in \mathbb{R}^l, l = 6\ to\ 10)$$

Car sales LSTM model

LSTM layer



LSTM cell



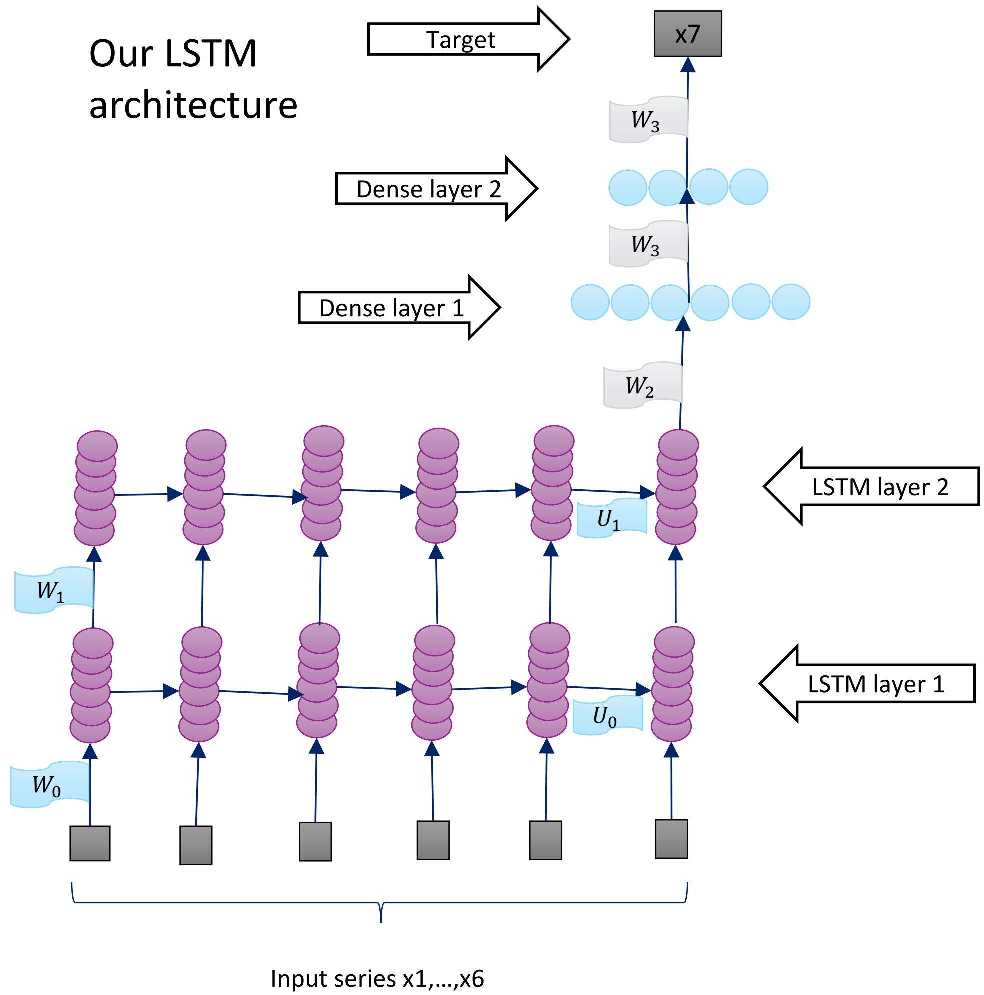
$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Our LSTM architecture



Performance valuation

- $y_t^p \ - true\ value\ of\ sales\ for\ PC\ p\ at\ time\ t$
- $\widehat{y_t^p} - our\ prediction\ for\ PC\ p\ at\ tme\ t$
- $err_t^p := \frac{\hat{y}_t^p - y_t^p}{y_t^p}$

- Baseline prediction is the naive (manager's) guess :

$$base\_err_t^p := \frac{y_{t-1}^p - y_t^p}{y_t^p}$$

- Compare histograms of err_t and $base\_err_t$(aggregate over PC's)