Humans and Algorithms: Creation and Measurement of Economic Value in Demand Forecasting

Peter Kauf, Prognosix AG
Thomas Ott, IAS, ZHAW
long shelf life
low costs
low margins
long shelf life
low costs
low margins

Short shelf life
high costs
high margins
Why forecasting?

food waste

0.7% - 3% of turnover

= 56.3 Bio CHF p.a. Loss

(NWS-Europe)
Why forecasting?

stock out

food waste

0.7% - 3% of turnover
= 56.3 Bio CHF p.a. Loss
(NWS-Europe)

1% - 2.3% of turnover
= 55.9 Bio CHF p.a. Lost turnover
(NWS-Europe)
We join Forces of Algorithms and People

PrognosiX
Comprehensive Forecasting

PrognosiX AG is a Spin-off from IAS Institute for Applied Simulation of ZHAW
Institute of Applied Simulation IAS
ZHAW Zurich University of Applied Sciences

- Bio-Inspired Modeling & Learning Systems
- Predictive Analytics
- Biomedical Simulation
- Applied Computational Genomics
- Simulation & Optimisation
- Knowledge Engineering

IAS:
- 6 research groups
- about 40 people
<table>
<thead>
<tr>
<th><strong>CTI project</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Denner</strong></td>
</tr>
<tr>
<td>Distribution center</td>
</tr>
<tr>
<td><strong>Migros Zürich, Fruits &amp; Vegetables</strong></td>
</tr>
<tr>
<td>Distribution center</td>
</tr>
<tr>
<td><strong>Bischofszell Nahrungsmittel</strong></td>
</tr>
<tr>
<td>Production Planning</td>
</tr>
<tr>
<td><strong>Inform Software</strong> <em>(Aachen, D.)</em></td>
</tr>
<tr>
<td>Demand planning add*ONE (Denner)</td>
</tr>
<tr>
<td><strong>Zürcher Hochschule für angewandte Wissenschaften</strong></td>
</tr>
<tr>
<td>Algorithms, interface and usability concepts</td>
</tr>
<tr>
<td><strong>PrognosiX AG</strong></td>
</tr>
<tr>
<td>Software development, commercialization</td>
</tr>
</tbody>
</table>
Learning algorithms
Add economic feedback

sales data

forecasting

algorithm
human overrides
error metrics

stock out, foodwaste, storage costs

external drivers
economic value
Add economic feedback

- sales data
- human expertise
- external drivers

- forecasting

- stock out, foodwaste, storage costs

- economic value

- human overrides
- error metrics

Library of algorithms
Add economic feedback

- sales data
- human expertise
- external drivers
- forecasting
- stock out, foodwaste, storage costs
- economic value
- human overrides
- error metrics
- Library of algorithms

Library of algorithms
Simple logic?

better forecasts

reduced leftovers / stockout

cost reduction

=> just pick the best forecasting method/algorithm
The goal of good forecasting is to minimize the forecasting error(s)

\[ e_t = F_t - X_t, \quad (1) \]

where \( X_t \) is the actual demand at time \( t \) and \( F_t \) is the respective forecast.

\( \Rightarrow \) How to quantify/evaluate the errors?

N.B. For now we assume that both \( X_t \) and \( F_t \) are available.
Overview:

- Standard accuracy measures / error metrics
- Advanced cost-based error metrics and sensitivity analysis
- Stock-keeping models
Measures of forecast accuracy

1. Scale-dependent metrics

The most popular measures are the mean absolute error (MAE)

\[ MAE(n) = \frac{1}{n} \sum_{t=1}^{n} |e_t| \]  \hspace{1cm} (2)

and the root mean square error (RMSE)

\[ RMSE(n) = \sqrt{\frac{1}{n} \sum_{t=1}^{n} e_t^2} \]  \hspace{1cm} (3)

Here and in the following we assume that the forecasting series is evaluated over a period \( t = 1, \ldots, n \).
How to choose the best algorithm? => Measures of forecast accuracy

2. Percentage error metrics

aim at scale-independence. E.g., the widely used mean absolute percentage error MAPE

$$MAPE(n) = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{e_t}{X_t} \right|$$

(4)
3. **Relative error metrics** compare the errors of the forecasting with the errors of some benchmark forecasting method. One of the measures used in this context is the relative mean absolute error (RelMAE), defined as

\[
\text{RelMAE}(n) = \frac{1}{n} \sum_{t=1}^{n} \frac{|e_t|}{|X_t - X_{t-1}|}
\]
4. **Scale free error metrics** have been introduced to counteract the problem “zeros in the denominators”. The **mean absolute scaled error** introduces a scaling by means of the MAE from the naïve forecast:

\[
MASE(n) = \frac{1}{n} \sum_{t=1}^{n} \left( \frac{|e_t|}{\frac{1}{n-1} \sum_{i=2}^{n} |X_i - X_{i-1}|} \right)
\]

(MASE (Hyndman and Koehler, 2006))
Measures of forecast accuracy

All measures come along with advantages and disadvantages

<table>
<thead>
<tr>
<th>Class</th>
<th>Advantage (e.g.)</th>
<th>Disadvantage (e.g.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale dependent metrics</td>
<td>Rather simple</td>
<td>No comparison across different time series</td>
</tr>
<tr>
<td>Percentage error metrics</td>
<td>comparison across different time series</td>
<td>Problems with small values /zeros in denominator</td>
</tr>
<tr>
<td>Relative error metrics</td>
<td>comparison across different time series</td>
<td>Problems with small values /zeros in denominator</td>
</tr>
<tr>
<td>Scale free error metrics</td>
<td>No problems with small errors</td>
<td>Interpretation of economic significance?</td>
</tr>
</tbody>
</table>

If we just want to know which is the best method—does it actually matter which metric to use?
Choosing the error metric

Yes, it matters sometimes!

Example: Sales sequence and two different forecasts for a convenience food product (both forecasting models based on regression trees)
Choosing the error metric

Which model should be chosen?
⇒ No coherent answer:

Peak model? Baseline model? Naive model?

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE</th>
<th>RMSE</th>
<th>MAPE</th>
<th>relMAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline model</td>
<td>2134.85</td>
<td>3234.45</td>
<td>0.1125</td>
<td>3.7248</td>
</tr>
<tr>
<td>Peak model</td>
<td>2080.34</td>
<td>2951.00</td>
<td>0.1253</td>
<td>6.2250</td>
</tr>
</tbody>
</table>

What model to choose? ⇒ What metric to choose?
How to decide?
Reasons for the differences?

• «Toy» example: Sales sequence (blue) with five disruptive peaks. A perfect baseline model (red) that misses the peaks and a perfect peak model (black) which is slightly shifted in between peaks.
Reasons for the differences?

• «Toy» example:

<table>
<thead>
<tr>
<th></th>
<th>MAE</th>
<th>RMSE</th>
<th>MAPE</th>
<th>relMAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline model</td>
<td>2.0</td>
<td>8.9</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>Peak model</td>
<td>0.95</td>
<td>0.97</td>
<td>0.10</td>
<td>2.74</td>
</tr>
</tbody>
</table>

• MAE/RMSE seem to put a heavier penalty on single high peaks than MAPE/relMAE

=> they favour the peak model over the baseline model

• Why so? We will see later
Economic significance of forecasting error

• The examples show an incoherent picture with regard to error metrics (which is also not remedied by the many alternatives that have been proposed in the literature)

• How to resolve the situation?

⇒ The actual core question is:
«What is the economic significance of the forecasts?»
I.e., «what are the consequences in terms of costs that come along with the forecasting errors?»
Cost-based error metrics

• Costs are product-specific and market-specific
• Real costs depend on many factors such as the stock-keeping process
• Simplest assumptions:
  – Forecast errors and costs are in direct relation
  – Costs do not depend on the history

\[ c \left( (X_t,F_t), (X_{t-1},F_{t-1}), (X_{t-2},F_{t-2}), ... \right) = c(e_t). \]  

(7)

• Example «ultra fresh products»
  – \( e_t > 0 \)  => forecast too high => foodwaste cost
  – \( e_t < 0 \)  => forecast too low => stock-out cost
Cost-based error metrics

- Generalised Mean Cost Error MCE (Ansatz):

\[
MCE(n) = s \left( \frac{1}{n} \sum_{t=1}^{n} c(e_t) \right)
\]

where \(c(\cdot)\) is a cost function and \(s(\cdot)\) is a scaling function.

- MAE and RMSE are special instances:
Cost-based error metrics

- **Linear MCE**: neglect economies of scale and assume proportionality:

\[
\text{linMCE} = \frac{1}{n} \sum_{t=1}^{n} c_{ab}(e_t) \quad \text{with} \quad c_{ab}(e_t) = \begin{cases} 
ae_t & \text{if } e_t \geq 0 \\
-be_t & \text{if } e_t < 0
\end{cases} \quad (9)
\]

*a*: cost per item for \( e_t > 0 \)
Cost per unsold item
=> foodwaste, storage

*b*: cost per item for \( e_t < 0 \)
Stockout cost
=> Non realised profit
Linear MCE: Sensitivity analysis

• For the example of «ultra fresh products»: linMCE expresses the cost due to foodwaste and stockout that results from forecasting errors

• In practice, it might be difficult to specify \( a \) and \( b \) for each product exactly.

  => make an **estimate** and perform a **sensitivity analysis** for a model comparison based on the **ratio** \( x = \frac{a}{b} \).
Linear MCE: Sensitivity analysis

Direct comparison of two forecasting models - Use ratio of linMCE:

\[
 f \left( x = \frac{a}{b} \right) = \frac{\text{linMCE}_{M1}}{\text{linMCE}_{M2}} = \frac{a \cdot \text{linMCE}_{a}^{M1} - b \cdot \text{linMCE}_{b}^{M1}}{a \cdot \text{linMCE}_{a}^{M2} - b \cdot \text{linMCE}_{b}^{M2}} \\
= \frac{x \cdot \text{linMCE}_{a}^{M1} - \text{linMCE}_{b}^{M1}}{x \cdot \text{linMCE}_{a}^{M2} - \text{linMCE}_{b}^{M2}} \quad (10)
\]

\[\text{linMCE}_{a}^{M_i}\] Sum of all positive errors for model \(M_i\)

\[\text{linMCE}_{b}^{M_i}\] Sum of all negative errors for model \(M_i\)
Linear MCE: Sensitivity analysis

«Toy» Example:

\[ f(x) \text{ can be determined analytically:} \]

\[
\frac{f\left(x = \frac{a}{b}\right)}{\text{linMCE}^{\text{baseline}}} = \frac{\text{linMCE}^{\text{peak}}}{0.95a} = \frac{2b}{x} = 2.11 \quad (11)
\]
«Toy» Example:

\[ f(x = \frac{a}{b}) = \frac{\text{linMCE}^{\text{baseline}}}{\text{linMCE}^{\text{peak}}} = \frac{2b}{0.95a} = \frac{2.11}{x} \] (11)

Conclusion:

The peak model performs better if the food-waste cost per item is smaller than 2.11 times the stock-out cost per item.

=> Baseline model with high stock-out costs during peaks
Linear MCE: Sensitivity analysis

• «Real world» example:

• Comparison against benchmark model (naive model)

\[ b_{\text{model}} \left( x = \frac{a}{b} \right) = \frac{\text{linMCE}_{\text{model}}}{\text{linMCE}_{\text{benchmark}}} \quad (12) \]
• «Real world» example:

1) $0 < x < 1.105$, the peak model outperforms the baseline model and the benchmark model, the benchmark model is the worst choice.

2) $1.105 < x < 2.050$, the baseline model outperforms the peak model and the benchmark model, the benchmark model is the worst choice.

3) $x > 2.050$, the baseline model is best, the peak model is worst.
Recapitulation: Cost-based error metrics

linMCE:

• Assumptions: Costs and errors in direct linear relation, no dependence on the history

• Estimate the economic consequences by estimating the parameters $a$ (cost per unsold item) and $b$ (non realised profit per item for stock-out)

• Perform a sensitivity analysis to assess the advantages of different forecasting models in dependence on the ratio $a/b$
Modelling „logistics“

Assumption: observed sales = simulated demand

Simplified ordering and stock keeping process

Week T:
stock at end of week $T = \max (\text{stock beginning of week } T - \text{demand in week } T, 0 )$

orders for week $T+1 = \max (\text{demand forecast for week } T+1 - \text{stock at end of week } T, 0 )$

Week $T+1$:
stock beginning of week $T+1 = \text{stock at end of week } T + \text{orders for week } T+1$

stock at end of week $T+1 = \max (\text{stock beginning of week } T+1 - \text{demand in week } T+1, 0 )$

orders for week $T+2 = \max (\text{Prognose Bedarf Woche } T+2 - \text{Bestand ende Woche } T+1, 0 )$

...  

Additional features:
  
  service level  ->  safety stock  
  shelf life     ->  product batches  
  delivery time  ->  transport logistics
10% stock keeping costs
20% margin
Service level 99%
## Results

10% stock keeping costs  
20% margin  
Service level 99%

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Baseline model</th>
<th>Peak model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average stock level</td>
<td>22’544 units</td>
<td>23’164 units</td>
</tr>
<tr>
<td>Safety stock level</td>
<td>3’886 units</td>
<td>3’415 units</td>
</tr>
<tr>
<td>Effective beta service level</td>
<td>98.08%</td>
<td>99.92%</td>
</tr>
<tr>
<td>Stock keeping costs</td>
<td>6’329 CHF</td>
<td>6’504 CHF</td>
</tr>
<tr>
<td>Opportunity costs</td>
<td>10’266 CHF</td>
<td>385 CHF</td>
</tr>
<tr>
<td>Stock keeping + opportunity costs</td>
<td>16’595 CHF</td>
<td>6’889 CHF</td>
</tr>
</tbody>
</table>
Comparison to linMCE

\[ a/b \approx 0.5 \]
Add economic feedback

- sales data
- human expertise
- external drivers

Forecasting

- stock out, food waste, storage costs

- human overrides
- error metrics

Library of algorithms

Economic value
„Essentially, all error metrics are wrong, but some are useful.“

(after George Box)