10 golden rules for applied data analysis (extended abstract)

Mats Carlin
University of Oslo
c/o SINTEF Electronics & Cybernetics

b/ot yaysU ot olcslCsC cr /sSy c/s Caca alaSntct cr CsdsSry tr4lCsU NrCsSSolT yUafcotst, -0 TrSCsl U4Sst aUs yUryrtsC O/of/ t/r4SC aSOant 3s vrSSrOsC, Baf/ nsaU a OoCs UalTs rv yaysUt ot y43Sot/sC O/of/ 3UsaRt trNs rv Nn yUryrtsC U4Sst,

The rule of problem knowledge. To become successful within the data analysis business, one must invest time and money in understanding the actual problem itself. Physical understanding, expert knowledge and explorative data analysis must be combined and every assumption must be checked to hold.

The rule of method knowledge. It is of vital importance that the data analyst knows the proper usage, benefits, assumptions and limitations of the method applied. Even the novice must have a basic understanding of standard statistical methods to succeed.

The rule of normalising data. Each variable should be scaled independently to the same range of values to avoid skewness in the estimation process and to assure robust and stable results.

The rule of operation domain. Data analysis models are only valid in those domains were the training data are dense enough for smooth model building. Whenever possible confidence intervals should be computed.

The rule of independent testing and benchmarking. Benchmarking should be done according to standard procedures and on several publicly available benchmark data sets. To make benchmark tests reproducible the test a
must be thoroughly describ

The rule of validatio
Whenever possible the dat
training, test and validation
test set may be used during
adaptation. The validation
to ensure objectivity.

The rule of fair comp
of time and effort should be
applied when performing a
easy to favour your own n
more time and effort in it
methods.

The rule of exposing
Whenever possible your da
edge about the data set sh
available to ensure reprodu
FTP or WWW are very su
data.

The rule of interpreta
analysis is interpretability
way we may gain new ge
the actual problem. When
choose models which are ea
pret.

The rule of scepticism
others methods and results
checked whenever possible.
field may be wrong some ti

I wish to acknowledge Y
University of Taiwan for he
suggestions of improvemen
stract would never had eme
ported by Hydro Aluminium
search Council (NFR) and