COMP 333 — Week 13 Welcome

Dangers in Data Analytics

In Week 13 we cover several recurring problems in Data Analytics

- Confusing correlation and causality
- Confounding variables
- Threats to validity
- Difficulties with Deployment
- Common Mistakes in Data Analysis

This material is to alert you to potential traps
in Data Analytics to do with the interpretation
of your models

and

the deployment of your models
with regard to

- maintenance
- scalabilty
- reproducibility.

Correlation is not Causality

Correlation is a relationship between two variables
where whenever one changes, the other is likely to also change.

Common Error This relationship, *correlation*, might lead us to assume
that a change to one thing **causes** the change in the other.

This is a conclusion we make all too often
So watch out for this bias in your own thinking.

The articles have many examples

where the correlation is obvious

and where the relationship is clearly not causality.

Explainability To demonstrate *causality*
you need to **explain** why it is a causal relationship.

This requires you to describe the *mechanism* of how changes in one variable

cause the second variable to change.

READ

Correlation is not causation

by Anthony Figueroa, 2019

`https://towardsdatascience.com/correlation-is-not-causation-ae05d03c1f53`

The second article presents experimental approaches
that might help distinguish causality from (mere) correlation.

Correlation vs Causation: Definition, Differences, and Examples
by Lionel Valdellon, 2019
`https://clevertap.com/blog/correlation-vs-causation/`

Confounding Variables

A *confounding variable* is a variable

that you did not collect or measure,

but which does play a role in your study,

so it should be in your model.

A *confounding variable* can have a hidden effect on the outcome of your analysis.

Think of it as the *"missing link"*
that would help you truly understand the data.

READ and, more importantly, WATCH the video

Confounding Variable: Simple Definition and Example
`https://www.statisticshowto.com/experimental-design/confounding-variable/`

Threats to Validity

Data analytics provides input to the decision-making processes
based on conclusions drawn from the models
that you have built from the data.

You need to ask: Are the conclusions valid?

Otherwise they will mislead the decision-making process,
and make wrong decisions with disastrous consequences!

So keep a checklist of what could go wrong,
what could threaten the validity of your conclusions.

This is particular important when your data concerns people,
and the decisions will affect people,
such as marketing, health, social media, government, etc.

Some Threats to Validity

- Confounding *(referred to above)*
- Historical events such as weekends, or non-trading days
- Testing participants repeatedly using the same measures
- Instrumentation may impact how participants respond
- Experimenter bias
- Situational factors such as time of day, location, noise
- Sample bias

Countermeasures Keep track of ways to counter, or remove, these threats
by thinking carefully about how you collect or select data.

READ

Understanding Internal and External Validity
by Arlin Cuncic, 2020
`https://www.verywellmind.com/internal-and-external-validity-4584479`

Issues in Deployment

There are several computing environments
in use for Exploratory Data Analysis

Commandline Scripting: such as the Unix shell with Unix tools
using text files, csv files for inputs, outputs, intermediate steps
with stepwise development of analysis
where script captures steps, parameters
Script is easy to replay

Notebooks: such as Jupyter that combines
interactive scripting with "literate programming"
that keeps track of thought processes during analysis
and allows analysis to be replayed

"Spreadsheets": such as OpenRefine, Excel, Tableau
that provide lots of tools and features, but little guidance
They need macros, and histories to capture/replay work
Most are proprietary, rather than open source

For deployment in production use, the environments are

IT infrastructure: traditional mix of DB servers, business workflow or web services,
and UIs for visualization and interaction
on desktop or mobile platforms.

Cloud: where all services are provded as SaaS (Software as a Service)
for data management, analysis software, UI software, and web services

Big Data infrastructure: provided by contracted service providers,
or pay-as-you-go through cloud SaaS, built using the Apache stack
with DB tools (HBase, HIVE)
and distributed computing tools (Hadoop, YARN, Spark)
supporting ML libraries (MLLib, GraphX)

Today, deployment across diverse platforms
is supported by virtualization and containers,
such as Docker and Singularity.

Best Practice

You should follow Software Engineering best practice.

However, you need to keep in mind the needs of ML systems

Build for reproducibility from the start: Persist all model inputs and outputs,
as well as all relevant metadata
such as config, dependencies, geography, timezones etc
Pay attention to versioning, including your training data.

Treat your ML steps as part of your build: automate training and model publishing

Plan for extensibility: If you are likely to be updating your models on a regular basis,
you need to think carefully about
how you will do this from the beginning.

Modularity: aim to reuse preprocessing and feature engineering code
from the research environment in the production environment.

Testing: Plan to spend significantly more time on testing your ML applications

Expand traditional SE testing to include:

Differential Tests: where you compare the average/per row predictions
given by a new model vs. the predictions
given by the old model on a standard test data set.

Benchmark Tests: to compare the time taken
to either train or serve predictions from your model
from one version to the next.

Load/Stress tests: are worth performing
given the unusually large CPU/Memory demand
of some ML applications

Reproducibility, Metadata, Regulation

You should aim for reproducibility from the start

If you work in a regulated environment,

such as healthcare and related research,

or finance and accounting

then detailed tracking of metadata and provenance

may be required by law.

READ

How to Deploy Machine Learning Models

by Christopher Samiullah, 2019

`https://christophergs.com/machine\%20learning/2019/03/17/how-to-deploy-machine-learning-`

Common Mistakes in Data Analysis

The recap concludes with future directions
for your learning about Data Analytics.

The associated article presents common and serious mistakes
that marketing studies make in Data Analytics.

These mistakes are issues, in general, in Data Analytics.

1. Vague Objectives
2. Sampling Bias
3. Unequal Comparison
4. Understanding the Meaning of a Metric
5. False Causality
6. Losing Sight of your Northern Star
7. Data Dredging
8. Getting Tunnel Vision
9. Lack of Statistical Significance
10. Relying on the Summary
11. Focusing on (or Ignoring) Outliers
12. Overfitting
13. Cherry Picking
14. Lacking Actionable Conclusions

READ

15 Most Common Deadly Mistakes in Data Analysis

by Mostafa El Bermawy, 2018

`https://nogood.io/2018/10/18/mistakes-data-analysis-marketing/`