Ethics of Data Science – Part II

A/B tests for Data Science and Machine Learning

A/B testing

Treatment

Randomly select 100 visitors to your website and show them a red button. Show a green button to everyone else.

Outcome

- *How many clicked?*
- *How many bought?*
- *How much time did they spend on the site?*

Statistical analysis

- *Binomial test, or Fisher exact test*
- *Student or Welsch t-test, Mann-Whitney U-test*

or logistic/linear regression

- *Did the success rate of the ad depend on the properties of the visitor (personalized advertising)? Build more complex regression models and look for interaction effects.*
- *Can we analyse this on retrospective, non-randomized data?*

Champion/challenger model testing

Although ML models are often used to extract insights about the effect of certain interventions, policies or decisions, increasingly they are production systems in their own right whose effectiveness we'd like to assess carefully.

In that case, the ML model **is** the treatment.

A=champion, B =challenger

Treatment

Randomly select 100 visitors to your website and show them ***the ad proposed by v2.0 of your recommender system.*** *Show the ad proposed by v1.0 to the rest.*

Outcome

- *How many clicked?*
- *How many bought?*
- *How much time did they spend on the site?*

Statistical analysis

- *Binomial test, or Fisher exact test*
- *Student or Welsch t-test, Mann-Whitney U-test*

or logistic/linear regression

When the model is a pure prediction system, and the "outcome" variable is its predictive accuracy, A/B testing it is the same as out-of-sample test performance, and does not need to be randomized. When however the model prediction is part of a larger business process, an A/B test is the golden standard.

Model monitoring and champion/challenger MLOps

Hot fix

Fix in production

Deterioriation-triggered

- *Monitor loss in production*
- *If increase past a threshold:*
  - *Retrain on fresh data*
  - *Rebuild from scratch*
  - *Rollback to last build*

Continuous testing

- *Permanently maintain an A/B setup with challenger*

Regular

- *Every x months, retrain or rebuild*
- *Run an A/B test*

- *More than one loss functions can and should be monitored, and a Pareto-like or AUC argument might be necessary to ascertain under what conditions challenger replaces champion.*
- *The technology to easily run A/B tests on models requires advanced MLOps.*

Model monitoring and champion/challenger MLOps

- Good Machine Learning Practices (GMLP)
- Patient-centric approach fostering transparency to users and trust
- Algorithm bias and robustness
- Real-world performance

- *"Algorithm Change Protocols"*
- Retraining strategies: what criteria must be met in production to trigger a more comprehensive performance evaluation.

Conclusion

- A/B testing is the business equivalent of randomized clinical trials
- It is also best practice when evaluating real-world performance of AI systems
- Understanding when/how to trigger a re-evaluation / re-training of an AI system is critical and relies on best practices around machine learning operations (MLOps).