Tuesday, 25 February 2020

Metrics in A/B testing

General division into:

1. General notes
2. Goal metrics
3. Driver metrics
4. Guardrail metrics
5. Other metrics
6. Notes

General notes:

- Might exist different metrics across teams / across company levels (e.g. company-level, team-level, feature-level, individual-level metrics)
- Each team is contributing differently to the overall success of the company
- Some teams are more focused on adoption, other on happiness, other on retention or performance
- Each team must articulate their goal and hypothesis on how their metrics relate to the overall company metrics

- The same metric may play a different role for different teams e.g. for Sales are latency and performance metrics as a guardrail, but for product development team those metrics might be goal metrics. On the other hand, product development have business metrics as guardrails.

Goal metrics

- Synonyms: success metrics / true north metrics
- One or a very small set of metrics
- Properties
    - simple: easily understood and broadly accepted
    - stable: not necessary to be updated every time you launch a new feature
- Should capture the ultimate success Avast is striving towards
- May not be easy to move in the short term
- Hint: try to define it in words what means “success” in Avast ($$$, engagement, retention, customer care, service, antivirus protection, etc.)
- What is Avast`s mission?
- Why is Avast antivirus successful as a product? What makes it successful?
- What does success looks like for Avast as a company?
- Answering these and similar questions should give us idea, what is the company direction and develop long term metrics such that go in hand with company evolution.

Driver metrics

- Short-term, faster-moving, more-sensitive metrics than goal metrics
- Properties
    - Aligned with the goal: it is important to validate that the drive metrics are in fact drivers of success
    - Actionable and relevant: teams must feel that they can act on features (product changes) to move these metrics
    - Sensitive: sensitive enough to measure impact from most initiatives
    - Resistant to gaming
- Popular frameworks:
    - The HEART framework:
        - Happiness, engagement, adoption, retention, task success

    - The PIRATE framework:
        - Acquisition, activation, retention, referral, revenue
- Frameworks can help to break down the steps that lead to success
- Good driver metric indicates that we are moving in the right direction to move the goal metrics

Guardrail metrics

- Protect the business and metrics that assess the trustworthiness and internal validity of experiment results
- Optimising goals and drives does not violate important constraints
    - Ex. 1: we want as many users as possible to register, but we do not want the per-user engagement level to drop drastically
    - Ex. 2: New feature vs. Load time
- Guardrail metrics are usually more sensitive than goal and driver metrics
- Sample Ratio Mismatch (SRM)
    - Should be included in EP such as guardrail metric
    - Chapter 21, page 182, Kohavi book
    - Bot detection: is it possible to exclude bots from the experiment? Mismatch might be caused by bots behaviour.
- Latency
    - Example from Bing, improving 1/10 sec is worth $18 million annual revenue
    - Testing latency: slow down experiment, Chapter 5, page 70, Kohavi book
    - Why Performance Matters (Wagner 2019)

Other metrics

- Data quality
    - Internal validity and trustworthiness of the underlying experiments
    - Are we tracking all screens properly?
    - Are we loosing some data? How much?

- Diagnosis / debug metrics
    - Helpful when debugging a scenario where the goal, driver or guardrail metrics indicate there is a problem

Notes

- CTR
    - CTR should be such as key metric
    - Should include other metrics indicating clicks on certain areas of the page e.g. click-through, close, timeout, time-spend and quick-backs, …
    - Penalise CTR for quick-backs where users come back quickly
    - Measure “success” (purchase), time-to-success is great sensitive
        - Click-through —> $$ (purchase induced by add)
- $$ (revenue)
    - Granularity:
        - Boolean 0/1: yes/no purchase
        - How much $
        - Average-order-value (AOV): average purchase value
- Renewal rate
    - One year subscription => year-long experiment (silly)
    - Surrogate metrics, e.g.: usage (measuring satisfaction)

- EP should have a few key metrics and hundreds to thousands of other metrics (segmentation), like in UWAM
    - segmentation by dimensions
        - Browsers
        - Markets
    - To make EP really useful it must contain all information available e.g. comparing clicks / $$ on country level. On the other hand users (e-comm team) will be forced to look up data themselves and the overall satisfaction / utility / benefit from building EP will be lost.

- One vs multiple metrics
    - multiple, cannot mix apples and pears
    - Counter example with jet airplane, chapter 7, page 90, Kohavi book
    - Tradeoffs between engagement and churn

- Overall Evaluation Criterion (OEC)
    - Idea of a single metric: (convex) combination of important business metrics

- Revenue generated from users clicking-through on the email
    - Nowadays we take into account all purchases! Does it make sense? Should not we track only those purchases, which preceded success click?

- more email => higher revenue
    - Spamming users
    - Treatment (email campaign) vs control (no email) => higher revenue
    - Penalisation?

- Email campaign
    - Penalisation unsubscribing!
    - Estimate “unsubscribe lifetime loss”, multiply it with number of users who unsubscribed and subtract it from revenue