Service Science
UniMiB
F9101Q022

Validated Learning

mirko.cesarini@unimib.it

Recap

- Main concepts learnt till now
  - Learning is the essential unit of progress for startups
  - Logic behind MVP building (Minimum Viable Product)
    - **Identify** the **riskiest assumptions**
    - Build an **MVP** to **test assumptions,** everything not related to learning is eliminated
    - **Iterate** the Build-Measure-Learn-Feed-back Loop to maximize the knowledge acquisition
- What will we do next?
  - How can we measure the learning achievements?
  - How to make adjustments to the service/product vision
  - How to prioritize where to invest in future development.

A Management History

- A manager was tasked to build a new product
- She/he goes back to her company's chief financial officer (CFO) after a year and says,
  - "We have **failed to meet** the **growth targets**. No new customers and no new revenue …"
  - "… However, we have **learned** an incredible amount and are on the cusp of a breakthrough new line of business. All we **need** is **another year**."
- Most of the time, this would be the manager last day of work in an organization

The importance of Measuring

- In general management, a **failure** to deliver results is due to
  - either a **failure** to **plan** adequately or …
  - … a **failure** to **execute** properly
- How can the CFO be sure that the manager is not lying i.e.,
  - the failure is not on execution …
  - … and the initial plan was wrong due to lack of knowledge?
- Call for **measuring** the learning achievements
  - If the manager can document learning achievements with empirical data … she/he won’t be fired (hopefully)
  - Management motto: if you can't measure it, you can't manage it

Validated Learning

- Goal: to empirically validate the learning achievements
- Learning is demonstrated (**validated**) by **metrics** positive **improvements**
  - If an MVP has no measurable metrics, it is not worth creating it.
    - it's easy to kid yourself about what you think customers want
    - It's also easy to learn things that are completely irrelevant
  - Suppose the Build-Measure-Learn loop was repeated several times
    - some service modification/improvement/… were tested, and measured
    - You can **detect from data** if there are some **improvements** (even small and limited)
    - e.g., # of monthly new customers (before & after the new service implementation)
- Challenge: to identify the **metrics** to measure the MVP achievements (and the start-up core activity performances in general)

Metric Example

- Suppose sales volume (can) measures a company growth
  - Sale profits are reinvested in marketing and promotions to gain new customers
  - The rate of growth depends primarily on
    1. the margin of each sale (sales - costs of goods sold)
    2. the repeat purchase rate of existing customers
    3. the cost of acquiring new customers
  - The higher 1. and 2. are, the lower 3. is, the faster the company will grow
  - The 1), 2), and 3) KPIs can be observed after each Build-Measure-Learn loop

Example

- Company selling a single product/service

	Months	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15
A	**Average Sale Price per unit**	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
B	**Average Cost per Unit**	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
C	**# New Users/Month**	4	2	3	4	3	4	5	18	32	56	131	262	786	1'965	4'912
D	**# Total Users**	4	6	9	13	16	20	25	43	75	131	262	524	1'310	3'275	8'187
E	**Avg # Items Purchased per User per Month**	1.0	1.0	1.0	0.8	0.8	0.7	0.9	1.1	1.3	1.6	1.8	2.4	2.8	2.9	3.0
F	**Avg # Items Purchased … considering only the last month new users**	1.0	1.0	1.0	0.5	0.5	0.5	1.5	1.5	1.5	2.0	2.0	3.0	3.0	3.0	3.0
G	**# (total) Purchases per Month**	4	6	9	7	8	10	38	65	113	262	524	1'572	3'930	9'825	24'561
H	**New Customer Acquisition Cost**	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
I	***Activities performed***				*Action A*			*Rollback A, Action B*			*Action C*		*Action D*			

- Considering the data shown here (A … H),
  Which is the best indicator(s) to evaluate the actions?

Considerations

	Months	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15
A	Average Sale Price per unit	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
B	Average Cost per Unit	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
C	# New Users/Month	4	2	3	4	3	4	5	18	32	56	131	262	786	1'965	4'912
D	# Total Users	4	6	9	13	16	20	25	43	75	131	262	524	1'310	3'275	8'187
E	Avg # Items Purchased per User per Month	1.0	1.0	1.0	0.8	0.8	0.7	0.9	1.1	1.3	1.6	1.8	2.4	2.8	2.9	3.0
F	Avg # Items Purchased … considering only the last month new users	1.0	1.0	1.0	0.5	0.5	0.5	1.5	1.5	1.5	2.0	2.0	3.0	3.0	3.0	3.0
G	# (total) Purchases per Month	4	6	9	7	8	10	38	65	113	262	524	1'572	3'930	9'825	24'561
H	New Customer Acquisition Cost	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
I	*Activities performed*				*Action A*			*Rollback A, Action B*			*Action C*		*Action D*			

- **F** shows the behavior of the last acquired customers in the last month
- **F** is better than **E** to highlight the impact change
- Rationale of focusing only on fresh customers
  - Good test-bed to assess changes
- **F** shows that learning was effective (**A** and **B** not considered, because they are constant in this case)
- Beware: **learning** has to be **validated** by **metric positive improvement**
- These just showed metrics are an example
  - there is no silver bullet i.e.,
  - different scenario may call for different metrics

Vanity vs Actionable Metrics

- E.g., in the previous case →
  - (D) and (G) are vanity metrics
  - (F) is an actionable metric
- **Actionable Metrics**:
  - allows people …
    - … to draw **cause-effect** inferences e.g., we are initially doing well because we are using all the revenue to buy new customers
    - … to identify problems. E.g., the initial (F) (A) (C) values cannot sustain long term customer acquisition
  - they help us …
    - … to design interventions e.g., to design Action A, B, and C
    - … and to evaluate the performances (Action A was bad, Action B was good)

A	Average Sale Price per unit
B	Average Cost per Unit
C	# New Users/Month
D	# Total Users
E	Avg # Items Purchased per User per Month
F	Avg # Items Purchased … considering only the last month new users
G	# (total) Purchases per Month
H	New Customer Acquisition Cost

Vanity vs Actionable … (2)

- E.g., in the previous case →
  - (D) and (G) are vanity metrics
  - (F) is an actionable metric
- **Vanity metrics**
  - They don't allow to draw cause-effect inferences , consequently, it is impossible to understand what is going on
    - E.g., Why the # of total users is increasing?
    - Is this sustainable?
  - Frequently they are (**too much**) **aggregated values**: a lot of aggregation prevent understanding the single driving forces/factors
  - The **curse of vanity metrics**: numbers look like good very frequently (D and G always increased) and make difficult to understand the underlaying behaviors / problems

A	Average Sale Price per unit
B	Average Cost per Unit
C	# New Users/Month
D	# Total Users
E	Avg # Items Purchased per User per Month
F	Avg # Items Purchased … considering only the last month new users
G	# (total) Purchases per Month
H	New Customer Acquisition Cost

3 A's of Good Metrics

- **Actionable** metrics help to draw **cause-effect inferences**, which help
  - to deeply understand what is going on (if you can understand …, you can wisely act)
  - to early identify problems
  - to design corrective actions
  - to evaluate action performances
- **Accessible** metrics
  - Can be easily **understood** by people
    - Keep it simple
    - Few indicators
  - Are easily **accessed** by the involved team, both in terms of
    - **Easiness** of access (e.g. reports sent weekly by email, web reports easily accessible)
    - **Evaluation criterion**: if people is not using the metric … there is a problem
- **Auditable** metrics can be easily verified
  - In case of bad results it is easy to blame "the messenger", the data quality, the data computation, … (bad results will come, for sure)
  - If everyone can check how metrics are computed, how data is retrieved, … this help reducing the blaming attitude

Metrics and Complexity

- The previous example was oversimplified
- Suppose a company sells several products with
  - Different margins
  - Different purchase rate behaviors
- Even metrics like (F), (A), and (B) may be not enough
- Call for in depth analysis
  - Cohort analysis
  - Split testing (for evaluating actions)
  - E.g., test a new product feature on a cohort and the old one on a different cohort

A	Average Sale Price per unit
B	Average Cost per Unit
C	# New Users/Month
D	# Total Users
E	Avg # Items Purchased per User per Month
F	Avg # Items Purchased … considering only the last month new users
G	# (total) Purchases per Month
H	New Customer Acquisition Cost

	Months	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15
A	Average Sale Price per unit	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
B	Average Cost per Unit	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
C	# New Users/Month	4	2	3	4	3	4	5	18	32	56	131	262	786	1'965	4'912
D	# Total Users	4	6	9	13	16	20	25	43	75	131	262	524	1'310	3'275	8'187
E	Avg # Items Purchased per User per Month	1.0	1.0	1.0	0.8	0.8	0.7	0.9	1.1	1.3	1.6	1.8	2.4	2.8	2.9	3.0
F	Avg # Items Purchased … considering only the last month new users	1.0	1.0	1.0	0.5	0.5	0.5	1.5	1.5	1.5	2.0	2.0	3.0	3.0	3.0	3.0
G	# (total) Purchases per Month	4	6	9	7	8	10	38	65	113	262	524	1'572	3'930	9'825	24'561
H	New Customer Acquisition Cost	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
I	*Activities performed*				*Action A*			*Rollback A,*			*Action C*		*Action D*			

Metric and Experiments

- You can use the identified metrics not only to test your initial idea …
- … but also, to run further experiments
  - E.g., to identify how to tune the growth engine
  - You can run experiments on client subsets
    - E.g., which is the best discount policy to use?
      - Better few under-the-costs items …
      - … or a lot of 20% discounted products?
  - Remember: you are not doing “in-silico simulations” (i.e., computer-based simulation in an artificial scenario)

Experiment Useful Methods

- Funnel analysis
- Cohort analysis
- Split Testing

Funnel Analysis

- Suppose you are monitoring how many of the people (who downloaded a game app application) have performed **In-App Purchases** (i.e. they get paid features)
- **Important steps** — **Conversion Rate**
  - Install App — 100%
  - Launch App — 90%
  - Reach Level 2 of Game — 45%
  - Offer Displayed — 45%
  - Purchase process started — 20%
  - Purchase Complete — 15%

- A funnel is a (people) flow in a sequential multi step process
- Conversion rate (for a specific step): how many people reach the step over the original ones
- Each funnel analysis is performed on a specific goal e.g.,
  - Signing up in a web site (i.e., splash page → demo → sign up)
  - # of software downloads in a web site
    (site visit → demographic data collection → disclaimer filling → Software download)

Cohort Analysis

- In statistics, a cohort is a group of subjects who share a defining characteristic
- Cohorts are useful
  - To breakdown people into subsets and analyze them separately e.g.,
    - Customers splitting based on the week they downloaded an app
    - Each group is a cohort
  - To perform experiments
    - E.g. several cohorts are selected (no overlapping people)
    - New product features are evaluated on the different cohorts

Split Test Experiments

- **Different** product **versions** offered to **several** customer **subsets** (e.g., cohorts)
- Insights can be discovered by observing group behavior differences
- This technique was pioneered by direct mail advertisers
  - Two versions of the same catalogue are sent to two groups of customers
    - The 2 catalogues have identical products but different design
    - 2 groups of different people with similar demographics were selected
  - The catalogue that lead the best performances (e.g., # orders) is the winner
- When the groups are two, this technique is called A/B testing

Reverse Order Planning

- MVPs and Experiments should be designed in reverse order as one might expect
  1. to figure out what is needed to learn
  2. to identify what should be measured to validate the learning
  3. to figure out how to build/improve the MVP
     to run the experiment and get the measurement
     - remark: an MVP is an early-stage product, not only a single test
     - Meaningful feed-backs can only be obtained from a complete product (even if at early stage), especially unforeseen issues

Experiment Example

- Scenario: **Banks** and the **Credit Card business** (U.S.A. - 1990)
- Important concepts (from the bank point of view)
  - Revenues: Card fees + interests
    - In the U.S.A. people can (and often do) reimburse expenses in several months
    - Interest are charged for reimbursements exceeding 1 month
  - Loss: customer defaults (bankruptcy)
- In (U.S.A. - 1990)
  - Uniform pricing and terms: everybody had the same **fees**, **credit limits**, **interest rates**
  - Bank competition focusing on
    - Enlarging the customer base
    - Avoiding customers having high default probability

Signet Bank Case

- Idea: Identify customers that will pay more on interests
- "Anyone can find customers who will take money and not pay you back! The trick is to find customers who will **take a lot of** your **money** fast and **pay** you **back slowly**."
  - Maximize the bank income on interests
  - Minimize the losses due to customer default
- Problem: banks did not know who were the "best profitable customers" according to the criteria above
- Idea:
  - Create different credit card products for several customer types
  - Build a predictive model to identify customer profitability (in addition to default probability), based on demography, age, income level, …
  - Identify the most profitable customers and solicit them to churn using marketing
- This idea was implemented at the Signet Bank (U.S.A.)

Data Problem - Experiments

- Which **characteristics** make different products **desirable** to **customers** and **profitable** for credit card issuers?
- **Not enough data** from the existing customer database
- Unfortunately, banks had collected data only …
  1. for the (single-type) terms offered in the past
  2. for customers who were deemed worthy of credit by the existing model
- **Experiments**: **different terms** were offered at **random** to **different customer cohorts**, then data was collected about customer behaviors
- Experiments focused not only on **riskier customers** but also on **less risky** ones
  - Nevertheless, the consequences of a lot of experiments was loosing money
  - In this case, **losses** are the **cost of data acquisition**
- Losses continued for a few years while the ~~data scientists~~ statisticians worked to build predictive models from the data, evaluate them, and deploy them

Signet - Success

- Finally, Signet's credit card operation turned around and became **so profitable** that the credit card business led to a **spin-off company**: Capital One (very famous in the U.S.)
- The new company grew to be one of the largest credit card issuers
  - Capital One acquired all the best customers
  - **Competitors** were left with the remaining ones, so, they **were forced to follow** or die
- In 2000, the bank reported to have carried out 45’000 of “scientific tests” as they called them
  - More details:
  - Clemons, Eric K., and Matt E. Thatcher. "Capital One: Exploiting an information-based strategy." hicss. IEEE, 1998.

Synthesis

- Experiment and learning goals: to find a synthesis between
  - The entrepreneur vision and …
  - … what customers would accept
- It is not
  - capitulating to what customers thought they wanted
  - to tell customers what they ought to want
- Validated learning: metric results will help going from opinions to facts