Product Sense -

Product/Feature Design & Launch Recommendations

Lecture Objective:

- Business sense towards A/B testing.
- How to launch A/B testing to get “valid” and “reliable” results.
    - Developing hypothesis
    - Designing A/B tests
    - Evaluating test results
    - Making decisions
- Launch recommendations and pitfalls with A/B testing set ups.

Example Cases: A/B testing Use Cases

- Multivariate Testing
- Split Testing
- Conversion Rate Optimization
- Landing Page Optimization
- Online Experimentations

Judgment Criteria & General Framework -

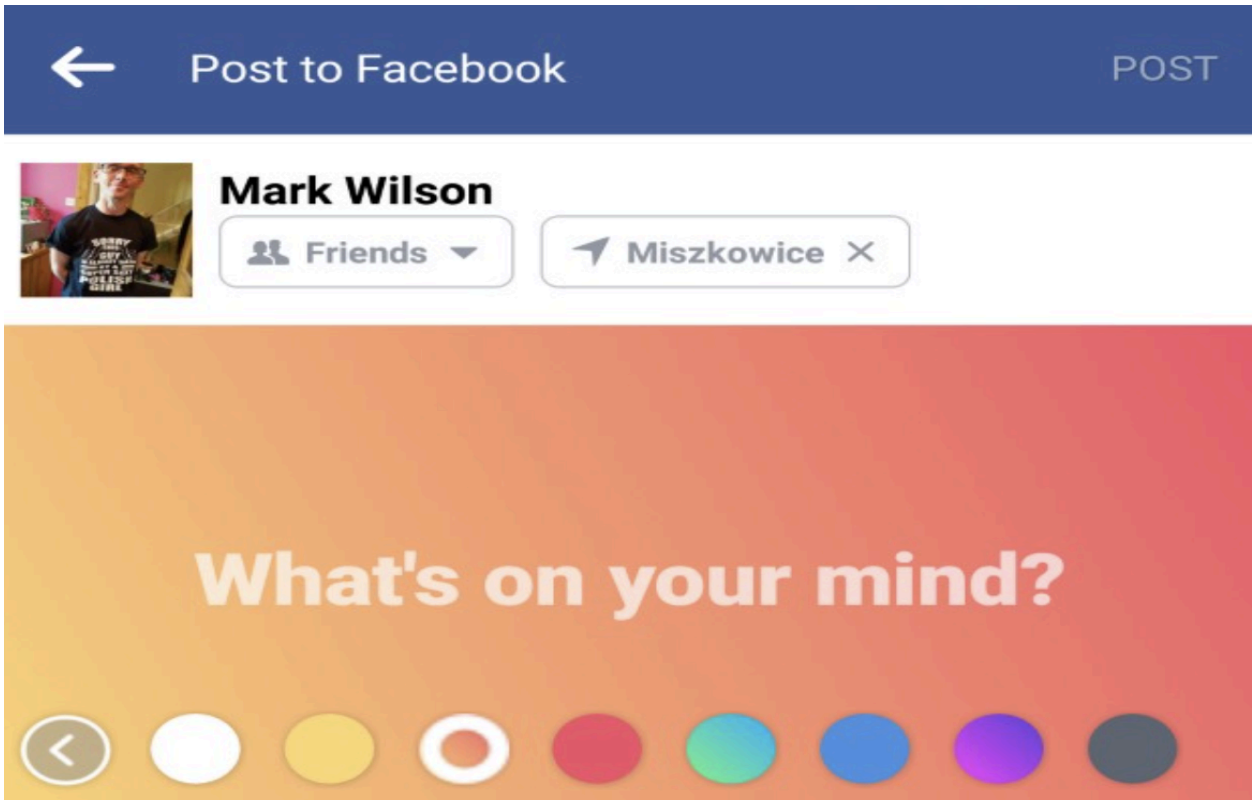
Keep this in mind when addressing business acumen questions.

- Design an experiment:
    - Structure
    - Right Flow of Steps
    - Develop the right Hypothesis

    - Sample Size Calculation
    - Duration
    - Pitfalls
- Identifying the right metric for testing significance - **North Star metric**
- Sample Segmentation for valid results - How would you take care of nuances and sample bias?
- How conclusively will the candidate be able to make a decision based on A/B test results?

Case Study:

How would you test if Facebook incorporating colored backgrounds to statuses improves user engagement?



1. **Clarify goal of the feature / idea conception :**

    a. What do you hope to achieve with this feature incorporation/update?
    b. Why this specific feature update for that goal and not any other feature.
    c. Has this been experimented before? Are other product lines also following suit?
    d. Is it because of previous experiment data, industry insights, reports or other evidence that supports your hypothesis?
    e. Is this feature for a specific user group or for all user groups?

2. **Discuss the metrics** that feature expects to bring an impact to and the data available.

    a. Give a list of metrics and finalize the metric we want to test significance on.
        i. Metrics - In this case study we can focus on the primary impact of the feature being more **user engagement**, **daily active users** etc.
        ii. **User Engagement** can be defined as % of active users who have engaged with Facebook in some way (like, comments, save, reactions).
        iii. **Daily Active Users** - # of unique users who have logged on Facebook each day. We expect this metric to increase with this new feature
    b. Come up with north **star metrics**, **supporting metrics** (if applicable) and **guard rail metrics.**
        i. **North Star Metric** - % of user with engagement
        ii. **Supporting Metric** - Daily Active Users
        iii. **Guard Rail Metric (**This is the metric that should not degrade in pursuit of a new feature**)** - % of media content (assuming media content provides more value, we don't want this % to decrease because of this feature). Or because this feature takes up so much space, are we seeing lesser number of posts on average that people interact with

3. **Experimentation** - How to design an experiment?

    a. Set up hypothesis (State the null hypothesis & alternate hypothesis) - What would be the null and alternative hypothesis in the case?
        i. Null Hypothesis - There is no significant difference in user engagement between the treatment and control groups.
        ii. Alternative Hypothesis - There is a significant difference in user engagement between the treatment and control groups.

    b. Choice of test - Since we are comparing two ratios, we can use the Z-proportions test.

    c. Choosing experiment control & treatment subjects
        i. Who is the experiment being run on?
        ii. Are we targeting all users on the platform? Or should we pick a proper segment of users for whom we feel this test will be particularly well suited

    d. Sample Size Calculation
        i. Baseline metrics
            1. Assume that before this feature launch, the user engagement is around 45%
        ii. Minimum detectable effect - what change is considered meaningful enough for you to take an action
            1. Assume that the business stakeholders are hoping for a 1% increase in user engagement in the treatment group
        iii. Significance level (Usually 95%)
        iv. Power (Usually 80%)
        v. With the above number, assume you would roughly need 40K users in each group for us to design this experiment in a statistically significant manner**.**

    e. Experiment Duration - Based on sample size estimated and the approximate traffic -

        i. Divide sample size by the number of users in each group
            1. Since we need a sample size of 80K (40K in each group) based on the above calculation
            2. Assuming FB gets a traffic of 5K everyday
            3. Experiment duration = 80/5 = 16 days
    f. Significance testing after we have reached the required sample size on the north star metric to identify significance.
    g. Continue monitoring supporting metrics and guard rail metrics.

4. **Testing Pitfalls -** How to avoid common challenges / experiment bias?
    a. **Experimental Design Bias**
        i. **Novelty / Primacy Effect** -
            1. **Primacy Effect -** When changes happen, some people that got used to how things work may feel reluctant to change
                a. Some users in the treatment group are reluctant to try out the new feature as they were used to the older status UI so they stop using FB much to post status.
                b. So user engagement for first 2 weeks are low Wk1 = 45% and Wk 2 = 48%
                c. But as these reluctant users see more users engaging with this colored status button, they will slowly start using this feature more.
                d. So from Wk3 onwards , the user engagement stabilizes to 62%
                e. It's important to not take the first 2 weeks of low user engagement due to the primacy effect into consideration when comparing with control.
                    i. Here it would have shown that there is no significant difference between the two groups in the first 2 weeks even though subsequently we see that this feature actually gets the users more engaged.

2. **Novelty Effect -** These users resonate with the new change and use more frequently
    a. Some users in the treatment group got excited about the new feature.
    b. The excited users use this feature and subsequently engage more in the first two weeks after which the excitement dies down.
    c. So user engagement for first 2 weeks are high Wk1 = 65% and Wk 2 = 68%
    d. But from Wk3 onwards , the user engagement stabilizes to 52%
    e. It's important to not take the first 2 weeks of high user engagement due to novelty effect into consideration when comparing with control.

3. Both of these effects are not long term effects, so it's important that results are not biased due to this effect.

    Treatment results may get exaggerated/undermined initially due to these effects.

4. Solutions:
    a. Run the experiment for a longer time than required if possible to observe for any novelty or primacy effect.
    b. The test can be conducted only on the first time users.
    c. Compare first time users with experienced users in the treatment group (we can get an estimated impact of primacy / novelty effect).

ii. Group Interference Qs -

Interference between variants happens a lot. It's important to select your sample in such a way that this interaction doesn't cause biased results.

1. Eg: IF the treatment group is seeing a positive effect because of this new FB status feature.
2. This effect can spill over to the control group (who is not seeing the new feature and makes new posts seeing their friend who is affected by the new feature in the treatment group). This is called a **network effect.**
3. So in this, the difference underestimates the treatment effect.
4. In reality the difference may actually be more than 1% but due to network effect, Actual Effect > Treatment Effect.
5. Hence giving an incorrect result that this new feature did not significantly impact the north star metric.

b. Outcome Bias

i. Look out for other design or system issues that led to the actual effect being undermined or over estimated to the treatment effect.

5. Recommendations based on experiment results - Launch or not?

a. Link results to the goal and business impact
   i. Example: What does 1% lift in engagement rate translate to revenue?
      1. If the 1% lift is increasing revenue through Ads by $20M, it might be worth it, however if it only increases revenue by $50K it might not be (based on efforts estimation).
   ii. Is it worth it to launch the product given all the costs?
   iii. While the perfect scenario is that the increase in success metrics are significant and we don't see any difference in the guardrail metrics - Give recommendations on what to do in case of conflicting situations.
   iv. Example: There's an increase in % user engagement among active users but also the daily active users have decreased.
   v. Translate this to impact to users and business -

1. Is the increased engagement among existing users bringing increased revenue to balance out the loss of some daily active users?
2. For eg:
    a. Let's say the daily active users were 5K earlier but now it has come down to 3K.
    b. However the user engagement has increased from 45% to 65%
    c. If the increase in user engagement has led to a revenue increase despite the loss of daily active users. This feature might be worth the consideration.
    d. ITs good to give a thought to strategy to retain the daily active users as next steps.

b. Consider short term and long term impact of the launch -
    i. Sometimes a short-term impression increase can conflict with the brand image or company's mission in the long run.
    ii. One reasonable suggestion could be even with the decrease in daily active users, the launch of color background search could potentially bring in more engaged users to the platform and in the long term, the benefit may outweigh the drawbacks.

Use Case of A/B Testing across various domains:

Instructor Note:

- Next we have some quizzes based on different use cases of A/B Testing.
- Launch the quiz and then ask the learners to share their views on goals of each of these, to make it more engaging and to help them understand better.

E-commerce Checkout Page Button Color:

Q. An e-commerce website wants to optimize its checkout page. They run an A/B test where half of the visitors see a green "Buy Now" button, while the other half sees a red "Buy Now" button.

What could be the goal of this experiment? Choose the most appropriate answer.

A. **The goal is to determine which color increases the click-through rate and leads to more purchases.**
B. The goal is to assess whether changing the button color impacts customer satisfaction with the website design.
C. The goal is to analyze the impact of button color on the average time spent on the checkout page.

Email Subject Lines for a Newsletter:

Q. A media company is testing different subject lines for its weekly newsletter. They send two versions of the newsletter, each with a different subject line, to a subset of their subscribers.

What would be the most appropriate action item for the media company as per the ongoing experiment?

A. They conduct a survey to gather feedback from subscribers about the subject line for their respective group.
B. They change the font style & color among the two versions of the newsletter to see if it has any impact on subscriber interaction.
C. They increase the frequency of newsletter distribution for one group to see if more frequent newsletters result in higher engagement.
D. **They measure open rates to determine which subject line is more effective at capturing reader attention.**

E-commerce Product Page Layout:

Q. An online retailer is experimenting with the layout of its product pages. They test two variations: one with the product description at the top and one with customer reviews at the top.

What would be the most appropriate choice of next step in the currently ongoing experiment?

A. They analyze the number of social media shares for each product to understand the impact of the layout variations on brand popularity.
B. **The company measures the click-through rate and conversion rate to determine which layout leads to more sales.**
C. They conduct a survey to gather customer preferences on the more appealing layout of the product pages of the website.

Streaming Service Content Recommendation Algorithm:

Q. A streaming service wants to enhance its content recommendation algorithm. They split their user base into two groups: one group receives recommendations from the current algorithm, and the other group experiences recommendations from a new algorithm.

What would be the most appropriate choice of next step in the currently ongoing experiment?

A. They analyze the number of products recommended during the testing period to assess the impact of the algorithm changes.
B. They monitor the recommendations generated by the new algorithm and compare them with the recommendations from their rival platforms.
C. **They compare user engagement metrics, such as time spent watching content and user ratings among the two algorithms to decide which one is to be kept as default for the entire user base.**

D. They analyze the variety of categories of products recommended during the testing period to assess the impact of the algorithm changes.
