A/B Testing Framework for Recommender Algorithms

We want to see which movie recommendation system (RS) works better at getting users interested in watching movies. Our main measurement is click-through rate (CTR) on recommended movies, which shows customers interest. We'll also look at how much they like the movies, how long they stay on the site, and if they keep coming back. We will use two-sample Z-test for difference between CTR proportions. If the Z-test p-value is less than 0.05 and the CTR for the treatment group is significantly higher than the control group, switch to the new algorithm.

Trafic split

We'll randomly assign users to two groups, one with the current algorithm (control) and one with the new algorithm (treatment). One group will use the current RS, and the other group will try the new one.

- Desired power: 80%
- Significance level: 5%
- Baseline CTR = 5%
- Desired increase = 1%

Using an online sample size calculator for comparing two proportions, we found that we need approximately 1570 users per group to detect this difference with the given power and significance level. With 4,000 users available, we can split them into two groups of 2,000 each, which is sufficient to meet the sample size requirement.

Process Overview

1. Pick the main things to measure to see if one RS is better than the other.
2. Randomly group the users, decide how many is needed for reliable results.
3. Use the chosen test to see which group clicks more on the movie recommendations.
4. Evaluate test results to decide if the new system is better and should be used from now on.

Example Walkthroughs

Testing a Content-based Algorithm

Test whether a content-based filtering algorithm increases user engagement compared to a simple collaborative filtering model.

- **Setup:** Users in the control group receive recommendations from a collaborative model, while those in the treatment group receive content-based recommendations using movie genres, release year, and user demographics.
- **Metrics:** CTR, average session length, user ratings of recommended movies.
- **Expected Outcome:** Improved CTR and user ratings for the content-based algorithm.

Testing a Hybrid Algorithm

Test whether combining collaborative and content-based features into a hybrid model results in higher satisfaction than the control.

- **Setup:** The control group continues with the current model, while the treatment group uses a new hybrid model that leverages user similarities and content features.
- **Metrics** CTR, retention rates, and user ratings.
- **Expected Outcome:** The hybrid model should ideally improve both the CTR and long-term retention.