A/B Testing Framework for a Book Crossing Site

Recommender Systems

July 2024

1 Objective

Our goal is to compare the implemented algorithms and determine which one provides the best book recommendations to users, leading to higher engagement and satisfaction.

2 Assumptions

- Users interact with the system by viewing and rating books and adding them to a reading list.
- System updates recommendations in real-time based on user interactions.
- System handles a number of users, substantial for statistically significant A/B testing.
- Continuous feedback from user interactions is used to improve recommendation algorithms.

3 Metrics

3.1 Primary

- **Click-Through Rate (CTR)**: The percentage of recommended books that are clicked by users.
- **Conversion Rate (CR)**: The percentage of recommended books that are rated or added to the user's reading list.

3.2 Secondary

- **Diversity of Recommendations**: Measure of how varied the recommended books are.
- **Algorithm Coverage**: The percentage of users and items for which the algorithm can generate recommendations.

4 Statistical Testing Approach

1. **Hypothesis**:
    - Null Hypothesis (H_0): There is no difference in performance between the two recommender algorithms.
    - Alternative Hypothesis (H_1): There is a significant difference in performance between the two recommender algorithms.
2. **Significance Level (α)**: 0.05 (5%)
3. **Statistical Tests**:
    - **Two-sample t-test** for comparing means of primary metrics between control and treatment groups.

5 Experiment Design

1. **Control/Treatment Split**:
    - We randomly assign users to either the control group (A) or the treatment group (B)
    - Ensure that the split is representative and unbiased
2. **Sample Size Calculation**:
    - Based on the desired effect size, statistical power, and significance level.
    - Sample size calculation for the two-sample t-test is made according to:

    $$n = \frac{2\sigma^2(Z_\beta + Z_{\alpha/2})^2}{\Delta^2}$$

    Where:
        - σ is the standard deviation of the metric
        - β is the z-score for the desired power
        - α is the z-score for the significance level
        - Δ is the minimum detectable effect size.
3. **Experiment Duration**:
    - Run the experiment for a sufficient period to gather enough data for statistical significance.
    - Monitor the metrics continuously to ensure the integrity of the experiment.

6 Validity Checks

- Ensure that between 2 algorithms there are no bugs and latency effects
- There were no unusual holidays or events during the test period
- Perform A/A test to ensure that the experimental setup, randomization process, and data collection methods are working correctly. So the differences in subsequent A/B tests are due to the changes in the treatment and not due to experimental bias or errors
- Check for a 'novelty effect' by segmenting old and new visitors
- Ensure that Chi Square test shows no significant difference between the user groups.

7 Decision-Making Methodology

7.0.1 Statistical perspective

1. **Analyze Results**:
   - Calculate the primary and secondary metrics for both control and treatment groups.
   - Perform statistical test to determine if the observed differences are statistically significant.
2. **Decision Criteria**:
   - If the p-value is less than the significance level (α), reject the null hypothesis and conclude that there is a significant difference between the algorithms.
   - Consider the practical significance of the results, not just statistical significance.
3. **Recommendations**:
   - If the new algorithm performs better on primary metrics and has no adverse effects on secondary metrics, consider deploying it system-wide.
   - If the results are inconclusive, consider running additional experiments or refining the algorithms.

7.0.2 Business perspective

1. Check for a practical usefulness of a degree of an effect. Whether improving the metric is good from business and ethical perspectives

2. Check for the metric trade-offs (how improving of one metrics affects another)

3. Calculate a cost-of-launch

8 Examples

8.1 Comparing Two Recommendation Algorithms

8.1.1 Setup

1. **Control Group**: Algorithm A (MF-based)

2. **Treatment Group**: Algorithm B (DNN-based)

3. **Metrics**: CTR (user clicks on a book), CR (user shelves a book)

4. **Sample Size Calculation**:

   - From a preliminary study we've got $\sigma = 0.05$ for CTR, $\sigma = 0.1$ for CR
   - Minimum detectable effect, $\Delta = 0.01$
   - For 80% power and 5% significance level:

   $$n = \frac{2 \times 0.05^2 \times (0.84 + 1.96)^2}{0.01^2} = 392$$

   $$n = \frac{2 \times 0.1^2 \times (0.84 + 1.96)^2}{0.01^2} = 1568$$

   - Each group should have at least 1568 users

5. **Experiment Duration**: 4 weeks, which is enough with a current service popularity to get sufficient sample size and compensate possible weekly and monthly trends. It is important not to interrupt the test, as intermediary results may be misleading.

6. **Statistical Test**: Two-sample t-test for CTR and CR

7. **Decision**: If p-value ¡ 0.05 and CTR, CR are higher for Algorithm B, deploy Algorithm B.

8.1.2 Simulation

Primary metrics Over a 4 weeks we've got total 9455 unique users, from which 4600 fall in A-group and 4855 into B-group. Running a chi-square test for SRM:

Chi-Square Statistic: 3.3855, p-value: 0.0658

	CR	CTR
A (MF)	0.05	0.1
B (DNN)	0.06	0.12

Table 1: Primary metrics

	CR	CTR
t-statistic	-5.56	-18.52
p-value	0.00	0.00

Table 2: T-test results

p-value is larger than 5%, so we assume that groups size mismatch is statistically insignificant. The sample size is enough for both CR and CTR metrics, which are as follows:

T-tests produce the following results: So we get a statistically significant difference between 2 algorithms for both metrics.

Secondary metrics

- Diversity of recommendations: both algorithms recommend the books based on ratings database and covers 99.99% of books in the dataset (1208 of 271379 are not rated) and from the existing knowledge perform equally
- Algorithm coverage: both algorithms are tested with a limitation for users with minimum 5 ratings

Business context Since the obvious improvement of recommendation quality and a fact that DNN-approach was already deployed for A/B testing it is reasonable to replace the current algorithm in favor of new one. Overhead of computational costs on a training process may be compensated by a better recommendations (and user satisfaction (to be measured separately).

8.2 Comparing Two Recommendation Algorithms

8.2.1 Setup

1. **Control Group**: Algorithm A (User-user collaborative filtering)
2. **Treatment Group**: Algorithm B (Item-item collaborative filtering)
3. **Metrics**: CTR (user clicks on a book), CR (user shelves a book)
4. **Sample Size Calculation**:
   - From a preliminary study we've got $\sigma = 0.1$ for CTR, $\sigma = 0.2$ for CR
   - Minimum detectable effect, $\Delta = 0.01$

- For 80% power and 5% significance level:

$$n = \frac{2 \times 0.1^2 \times (0.84 + 1.96)^2}{0.01^2} = 1568$$

$$n = \frac{2 \times 0.3^2 \times (0.84 + 1.96)^2}{0.01^2} = 6272$$

- Each group should have at least 6272 users

5. **Experiment Duration**: 4 weeks, which might be enough with a current service popularity to get sufficient sample size and compensate possible weekly and monthly trends. It is important not to interrupt the test, as intermediary results may be misleading.

6. **Statistical Test**: Two-sample t-test for CTR and CR

7. **Decision**: If p-value ¡ 0.05 and CTR, CR are higher for Algorithm B, deploy Algorithm B.

8.2.2 Simulation

Primary metrics Over a 4 weeks we've got total 8150 unique users, from which 3700 fall in A-group and 4450 into B-group. Sample size is not sufficient for conducting tests for CR.

Running a chi-square test for SRM:

Chi-Square Statistic: 34.3982, P-value: 0.0000

p-value is 0%, so the group size mismatch is statistically significant. Which points into problems in randomization process.

Conclusion Revise the group splitting algorithm, perform test for a longer period (2 months) to get sufficient number of user interactions

8.3 Comparing Two Recommendation Algorithms

8.3.1 Setup

1. **Control Group**: Algorithm A (User-user collaborative filtering)

2. **Treatment Group**: Algorithm B (Item-item collaborative filtering)

3. **Metrics**: CTR (user clicks on a book), CR (user shelves a book)

4. **Sample Size Calculation**:

   - From a preliminary study we've got $\sigma = 0.1$ for CTR, $\sigma = 0.2$ for CR
   - Minimum detectable effect, $\Delta = 0.01$

	CR	CTR
A (UU)	0.081	0.0701
B (II)	0.085	0.0730

Table 3: Primary metrics

	CR	CTR
t-statistic	-1.4426	0.069
p-value	-1.8183	0.1492

Table 4: T-test results

- For 80% power and 5% significance level:

$$n = \frac{2 \times 0.1^2 \times (0.84 + 1.96)^2}{0.01^2} = 1568$$

$$n = \frac{2 \times 0.3^2 \times (0.84 + 1.96)^2}{0.01^2} = 6272$$

- Each group should have at least 6272 users

5. **Experiment Duration**: 8 weeks, which might be enough with a current service popularity to get sufficient sample size and compensate possible weekly and monthly trends. It is important not to interrupt the test, as intermediary results may be misleading.

6. **Statistical Test**: Two-sample t-test for CTR and CR

7. **Decision**: If p-value ¡ 0.05 and CTR, CR are higher for Algorithm B, deploy Algorithm B.

8.3.2 Simulation

Primary metrics Over a 8 weeks we've got total 15986 unique users, from which 7838 fall in A-group and 8058 into B-group. Sample size is not sufficient for conducting tests for CR.

Running a chi-square test for SRM:

Chi-Square Statistic: 1.4949, P-value: 0.2215

p-value is much larger than 5%, so we assume that groups size mismatch is statistically insignificant. The sample size is enough for both CR and CTR metrics, which are as follows:

T-tests produce the following results: So we can not reject the null-hypothesis for both metrics and can't conclude that algorithm B is more efficient.

Conclusion

- There's no evidence that one of the algorithms is better than the other.

Business context There's no obvious argument for favoring A or B approach. So, we'll not redesign the system and leave it intact.