A/B Testing Framework
Objective
Outline the methodology for A/B testing of recommender algorithms on your dataset. You should make assumptions about the online system characteristics in case
Deliverables
Markdown or pdf file with a defined framework for comparing recommender algorithms on your dataset. This document should go over metrics selection, statistical testing approach, experiment design (control/treatment split, sample size calculation based on the effect size, statistical power, and significance level), and the methodology for decision-making. In addition to the framework definition, you should provide at least two detailed walkthrough examples.

As we know, the main task of any A/B testing is to test a certain hypothesis. Since this course is called recommender systems, our hypotheses should be related to them. The two main hypotheses that can affect our recommender system are:

1) Does the new feature have a sufficiently strong influence on the outcome of the model.

2) Check whether one model is better than another.

In the next steps we will build a frameworks for both this hypothesis and check if we should change our model, or previous model is also good enough.

Example 1

Description

Let's take a first example. Let's say we decide to add an analysis of user comments using ChatGPT as a new variable to our model. Most likely, such a step can greatly improve the accuracy of our recommender system, but in this case we also need to understand that this model is not free and when comparing, we need to consider the price that it will cost us to add this variable to our model.

Prerequisites

Metric Selection

Obviously, the main metric that any business wants to raise is Revenue, but in our case, we don't have any data that can help calculate this metric, so a better option would be to use another metric.

One of the best metrics that can help is the return rate. This variable indicates the percentage of users who will visit another restaurant in a given period of time.

In general, we can divide our customers into two groups:
1) Those who will come again regardless of our recommendation system.

2) Those who came through our recommendation.

With a sufficiently large number of users, we can say that first group will be equal in both tests, so the evaluation of systems using this metric is a correct example.

Division

Another important criterion is the correct distribution of our groups. In order to make sure that our groups are similar, we can divide our users by the number of visits to our establishments so far. This distribution will help us even more to neutralize the influence of users who visit our establishments constantly independently of the recommendation system and focus on a second group: Those who came through our recommendation.

Experiment Design

First of all, we need to define a test group, it is obvious that since our metric is Return Rate, the best option for the group are customers who visited one of our establishments. Since our data only consists of users who have visited and rated our establishment, we can use our entire dataset for testing.

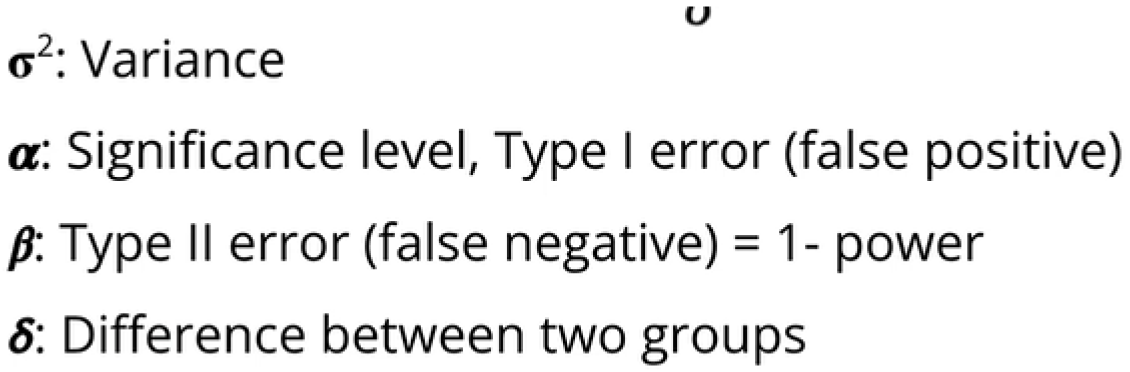
Let's assume that after calculations, our company came to the conclusion that in order to add the new feature built on ChatGPT, the Return Rate needs to increase by 1%, this will be our Effect Size

Then we will use Significance Level = 0.05 and Power = 0.8 which is the industry standard.

Now we could calculate size using following formula:

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

Where:



$\boldsymbol{\sigma}^2$: Variance

$\boldsymbol{\alpha}$: Significance level, Type I error (false positive)

$\boldsymbol{\beta}$: Type II error (false negative) = 1- power

$\boldsymbol{\delta}$: Difference between two groups

We could also calculate $z_{a/2}$ and z_B

$\alpha = 0.05$, then $z_{\alpha/2} = z_{.025} \approx 1.96$
$\beta = 0.2$ (power $= 0.8$), then $z_{\beta} = z_{.20} \approx 0.84$

$$n = \frac{2\sigma^2 * 2.8^2}{\delta^2} = \frac{16{*}\sigma^2}{\delta^2}$$

We could not calculate variance and difference in our groups using our data, so let's assume that the test size should be equal to 1000.

Methodology for decision-making

Null Hypothesis (H0): The mean return rate of the current system is greater than or equal to the new system ($RR_A + d < RR_B$).
Alternative Hypothesis (H1): The mean return rate of the current system is less than the new system ($RR_A + d > RR_B$)
Then we could use a one-sided Z-test to check the hypothesis.

As we couldn't calculate effect, let's assume some resulting value

$RR_A = 0.6$
$RR_B = 0.585$
$N_A = N_B = 1000$
d = 0.01

We also know that for each individual user we have two states:
1) He visited another restaurant
2) He did not visit another restaurant.

From this we can conclude that our distributions in both cases are Bernoulli distributions. So now it is also very easy for us to calculate the std.

$VAR_A = 0.6 * 0.4 = 0.24$
$VAR_B = 0.585 * 0.415 = 0.242775$

As out N >> 30 we could use Z-test.

$$Z = \frac{\bar{x}_B - \bar{x}_A}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

Z = (0.6 - 0.585) / sqrt(0.24/1000 + 0.242775/1000) = 0.227560887

For a one-tailed test at α=0.05, the critical value is $Z_{0.05}$≈1.645 so we couldn’t say that new model is better.

Example 2

Description

In this case we want to compare two models. We decide that both models are similar in price.

Experiment Design

Almost everything is identical to the first option, except that since the models are the same price, the second model should simply be better than the first.

Methodology for decision-making

Null Hypothesis (H0): The mean return rate of the current system is greater than or equal to the new system ($RR_A < RR_B$).
Alternative Hypothesis (H1): The mean return rate of the current system is less than the new system ($RR_A > RR_B$)
Then we could use Z-test to check hypothesis.

As we couldn’t calculate affect let’s imagine some value

RR_A = 0.625
RR_B = 0.585
$N_A = N_B$ = 1000

VAR_A = 0.625 * 0.375 = 0.234375
VAR_B = 0.585 * 0.415 = 0.242775

$$Z = \frac{\bar{x}_B - \bar{x}_A}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

Z = (0.625-0.585) / sqrt(0.234375 + 0.242775) * sqrt(1000) = 1.83 > 1.645

We could say that new model is better, and we should change our model.