Experiment Design for Optimization

Sonal Shad, David Ramirez, Cassie Richter, Inseong Han

Contents:

Challenge

Created by N. Stevens.

This challenge is inspired by the culture of experimentation at Netflix with a hypothetical problem and a web-based response surface simulator. Our goal is to optimize the Netflix homepage by way of *minimizing browsing time*.

Problem: When faced with so many viewing options, Netflix users often experience choice-overload and can be overcome by decision paralysis, which negatively impacts Netflix because a user may become overwhelmed and may ultimately lose interest and not watch anything. We conduct a series of experiments to learn *what* and *how* the following 4 factors influence browsing time:

- Tile Size: Ratio of a tile's height to the overall screen height. Aspect ratio is fixed
- Match Score: Predicted likelihood of user's enjoyment.
- Preview Length: the duration (in seconds) of a show or movie's preview.
- Preview Type: The type of preview that autoplays (Teaser/Trailer vs Actual Content)

Experimentation:
We access the response surface simulator on the web. We upload a design matrix (experimental conditions) and collect results. The simulator mimics the random assignment of n=100 users to each condition and returns the response variable, i.e. browsing time. We are limited to a maximum of 40 experiments, which mimics real-life budget and time constraints.

Executive Summary

This project aimed to minimize the amount of time users spend browsing Netflix to mitigate loss of user engagement due to decision paralysis. e examined the influence of 4 layout factors on the platform to find the optimal configuration that minimized the average browsing time. We randomized 100 users to each of several experimental conditions (with varying factors *Preview Length, Preview Type, Match Score,* and *Tile Size*). We identified the optimal condition with *Preview Type:* Teaser/Trailer, *Preview Length*: 75 secs, *Match Score*: 74, and *Tile Size*: 0.2 to have a mean browsing time of 10.099 mins.

Introduction

Many users grapple with choosing what to watch from the vast array of TV shows and movies available on the Netflix platform, leading to prolonged browsing. We seek to reduce this browsing time so that users are less likely to face decision paralysis and disengage from the platform. We attempt to find the optimal levels of 4 components of the homepage layout, our ‘design factors’ within their region of operability (Table 1).

Design Factor	Description	Region of Operability	Default Value
Match Score	Integer value for match score percentage predicting user enjoyment	[0,100]	95
Preview Length	Length of preview shown. Can only be set in 5 second intervals.	[30,120]	75
Preview Type	Whether preview shown is a Teaser Trailer or Actual Content	TT or AC	TT
Tile Size	Ratio of tile height to screen height	[0.1, 0.5]	0.2

Table 1. Design factors and their possible levels.

We adopt a serial experimentation strategy, beginning with a 2^4 factorial experiment to identify factors and interactions that significantly influence the average browsing time. In the second phase, we optimize the values of the significant factors by experimenting with additional levels of each, and employing techniques such as response surface methodology to narrow down the range of levels. In the end, we identify a smaller range of optimum levels and collect data in the region to empirically find the condition with the minimum average browsing time.

Experimentation

In our experiments, we randomized 100 Netflix users to experimental conditions with unique combinations of the factors' chosen levels and recorded their browsing time. A sample size of 100 per condition was assumed to be sufficient without power analysis. The results of all hypothesis tests were evaluated at a significance level of 0.05.

Factor Screening

Round 1

We screened the factors by performing a 2^4 factorial experiment i.e. with 16 experimental conditions. *Preview Type* was tested at both TT, AC conditions. The levels of the numeric factors were chosen to be distinguishable from each other while avoiding extreme values in their operational range. Care was also taken to choose values proximal to the default values. The selected levels were *Preview Length* (55, 95), *Match Score* (50, 80), and *Tile Size* (0.2, 0.4).

We fit a full linear regression model on the collected data with terms for the main effect of each factor, as well as all the possible interaction terms (Model 1).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_3 x_3 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3 + \beta_4 x_4 + \beta_{14} x_1 x_4 + \beta_{24} x_2 x_4 + \beta_{124} x_1 x_2 x_4 + \beta_{34} x_3 x_4 + \beta_{234} x_2 x_3 x_4 + \beta_{1234} x_1 x_2 x_3 x_4$$

$$Model\ 1.\ y = mean\ browsing\ time,\ x_1 = Preview\ Length,\ x_2 = Match\ Score,\ x_3 = Preview\ Type\ and\ x_4 = Tile\ Size.$$

Following the analysis of t-tests, we ascertained that *Preview Length, Preview Type,* and *Match Score* have significant main effects on average browsing time, while *Tile Size* does not. Additionally, the tests indicate the significance of a singular interaction effect, specifically between *Preview Length* and *Match Score* (Table 2).

We created a reduced model containing only the significant main effects and interactions (Model 2), compared it against Model 1 through a partial F test, and concluded that there was no loss of information between the two models (p-value = 0.572).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2$$

$$Model\ 2.\ y = mean\ browsing\ time,\ x_1 = Preview\ Length,\ x_2 = Match\ Score,\ x_3 = Preview\ Type$$

In the optimization phase, we conduct further tests to determine the optimal values of *Preview Length, Preview Type*, and *Match Score* with *Tile Size* at its default value of 0.2.

	coef	std err	t	P>|t|
Intercept	17.3133	0.025	704.486	0.000
x1	0.1768	0.025	7.193	0.000
x2	-1.5956	0.025	-64.928	0.000
x1:x2	1.2133	0.025	49.369	0.000
x3	2.4808	0.025	100.946	0.000
x1:x3	0.0448	0.025	1.823	0.068
x2:x3	-0.0086	0.025	-0.350	0.727
x1:x2:x3	0.0422	0.025	1.718	0.086
x4	0.0173	0.025	0.705	0.481
x1:x4	0.0032	0.025	0.130	0.896
x2:x4	-0.0077	0.025	-0.315	0.753
x1:x2:x4	-0.0110	0.025	-0.446	0.655
x3:x4	-0.0203	0.025	-0.824	0.410
x1:x3:x4	-0.0062	0.025	-0.254	0.799
x2:x3:x4	0.0220	0.025	0.895	0.371
x1:x2:x3:x4	0.0208	0.025	0.846	0.398

Table 2. y ~ x1 * x2 * x3 * x4

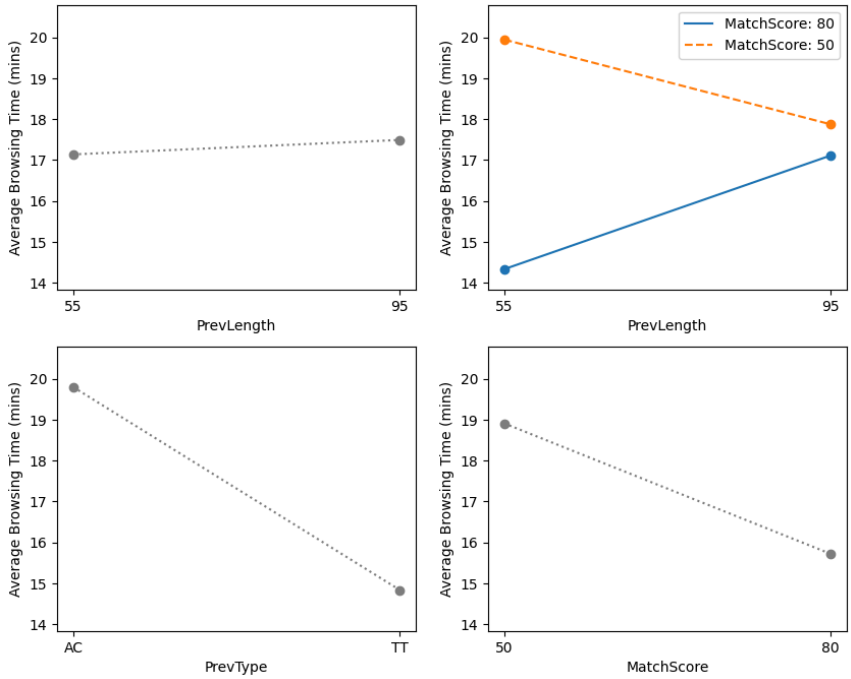

Figure 1. Significant Main and Interaction Effects

Optimization

Preview Type

Model 2 indicates that *Preview Type* has a notable effect on average browsing time, particularly increasing it by 2.4808 mins when the level is 'Actual Content', compared to 'Teaser/Trailer' (Table 2). Since it does not interact with other factors, we determined that 'Teaser/Trailer' is the optimal *Preview Type*.

To streamline further experimentation, we excluded *Preview Type* from our new model (Model 3) and focused our experiments on data (both existing and new) with the TT *Preview Type.*
We confirmed that there is no loss of information in using Model 3 (on data with only TT *Preview Type*) by comparing it with Model 2 using a partial F-test (p-value = 0.3223).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_1 x_2.$$

$$Model\ 3.\ y = mean\ browsing\ time,\ x_1 = Preview\ Length,\ x_2 = Match\ Score$$

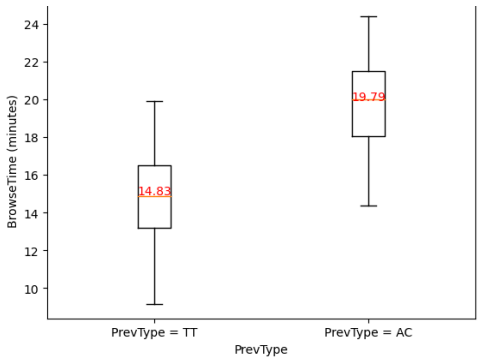

Figure 2. Comparison of Mean Browsing Time with different levels of Preview Types

Preview Length & Match Score

Subsequently, we focused on optimizing the combination of *Preview Length* and *Match Score*. We collected data through multiple rounds of experimentation, introducing new levels of the two factors. Importantly, in this simulation, external factors such as time do not influence browsing time, allowing us to pool new and previously collected data without confounding results.

Round 2

The factorial experiment in Round 1 revealed that browsing times were lower with *Match Score* at 80, compared to 50, and *Preview Length* at 55 than 90 (seconds). We tested 4 additional conditions, incorporating combinations of *Match Score* of 65 and 90, and *Preview Length* of 45 and 75. With data from 12 experimental conditions (restricting conditions from the first round to ones with *Preview Type* TT), we construct a second-order model (Model 4) with quadratic effect terms in addition to the terms in Model 3. All terms in this model were significant (p-value = 0.000) which indicated that our selected region was in the general vicinity of a minimum.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} {x_1}^2 + \beta_{22} {x_2}^2$$

$Model\ 4.\ Second\ Order\ Quadratic\ Model\ with\ significant\ main\ and\ interaction\ terms$

$y = mean\ browsing\ time,\ x_1 = Preview\ Length,\ x_2 = Match\ Score$

Using Model 4 and data from 2 rounds of experiments, we estimated a response surface and the corresponding stationary point (*Preview Length* 69.59, *Match Score* 73.76, estimated average browsing time 10.822 mins) in the region of operability (Figure 3).

Round 3

For the next round of experimentation, we defined 9 new experimental conditions using a Central Composite Design with the center at the closest practical value to the previously estimated stationary point (*Match Score* 74, *Preview Length* 70).

Our factorial design points were at a distance of 4 natural units from the central *Match Score* on either side, and 5 natural units from the central *Preview Length* on either side. The 4 axial points were at a distance of a =$\sqrt{K'}$=$\sqrt{2}$ = 1.414 coded units from the center. These are visualized in Figure 3.

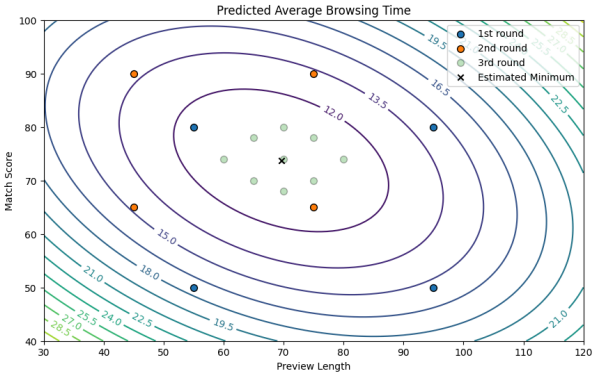

Figure 3. Contour plot of the RS with expected minimum after and 2 rounds and experimental conditions from Round 3.

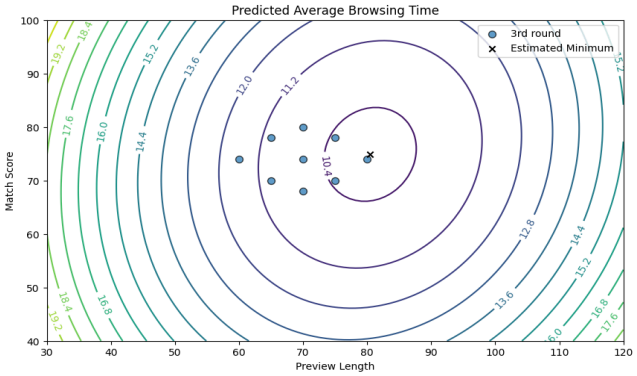

Figure 4. Contour plot of RS with calculated optimum using Round 3 data

After fitting Model 4 on only the data from Round 3 of experiments, we create a second, more condensed response surface (Figure 4). Figure 4 also highlights the stationary point on the newly estimated response surface (*Preview Length* 80.47*, Match Score* 72.95*,* estimated average browsing time 10.236 mins).

Rounds 4+

Using this predicted optimum from the response surface and the observed mean browsing times (Table 3, Figure 5) to narrow our range of factor levels, we conduct several experiments with various new conditions in the vicinity to empirically determine the condition with the minimum average browsing time (10.099 mins at *Preview Length* 75 *Match Score* 74).

		Mean Browsing Time
Preview Length	**Match Score**	
60	74	11.563947
65	70	10.990166
	78	11.057552
70	68	10.683007
	73	10.521104
	74	10.619180
	80	10.691765
75	70	10.306197
	73	10.190853
	74	10.099106
	75	10.185624
	78	10.322490
80	73	10.329073
	74	10.273886
	75	10.186486
	76	10.646209
	77	10.628726

Table 3. Mean Browsing Time (mins) at different Match Score and Preview Length level

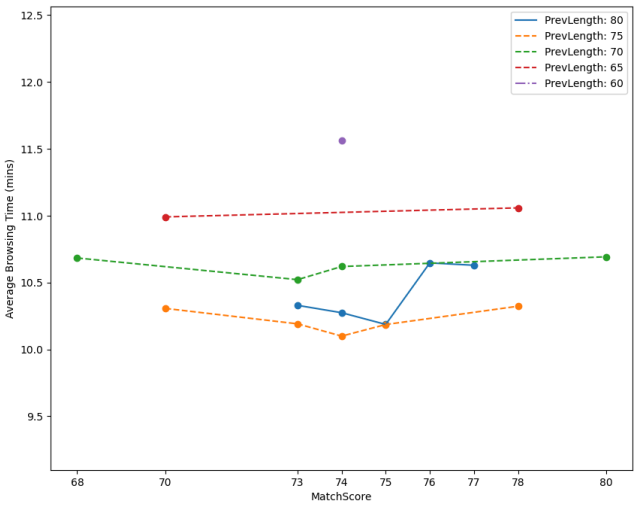

Figure 5. Interaction plot of Mean Browsing Time at Levels of Preview Length and Match Score with Experimental Conditions from Round 3 onwards

Conclusion

After conducting experiments with 4 factors that we presumed to influence browsing times, we discovered that *Preview Length*, *Match Score*, and *Preview Type* are important factors, whereas *Tile Size* is not. Additionally, we identified a significant interaction between *Preview Length* and *Match Score*.

Overall, we tested 37 experimental conditions with 100 samples each, to find the following optimum condition:

Match Score	Preview Length (s)	Preview Type	Tile Size	Average Browsing Time (mins) [95% Confidence Interval]
74	**75**	**Teaser/Trailer**	**0.2**	**10.099** [9.901, 10.297]

One limitation of this study is that we only tested 4 factors, 1 of which was insignificant. In reality, there are many factors that would influence browsing time, controllable as well as uncontrollable factors that would need to be considered to ensure optimum conditions. Another limitation is the possibility that we made a type II error when eliminating the *Tile Size* factor, and/or choosing to hold *Preview Type* constant, missing significant main and/or interaction effects of these factors. Lastly, since the second order model is better at finding a minimum in a smaller, localized region, we may have found a local minimum if our initial search region was actually far from the global minimum.