Netflix Homepage Optimization

Experiments in Data Science - MSDS 629

Chris Nishimura

I. Executive Summary

In this series of experiments, the goal was to optimize the Netflix homepage in order to minimize a user's average browsing time. In this Netflix-inspired project, the factors I was able to manipulate and their ranges were **`Tile.Size`** (as proportion of page height) $[0.1, 0.5]$, **`Match.Score`** $[0, 100]$, **`Prev.Length`** (preview length in seconds) $[30, 120]$, and **`Prev.Type`** (preview type) $\{Teaser/Trailer\ (TT),\ Actual\ Content\ (AC)\}$. I conducted a series of experiments, a 2^k factorial experiment followed by two central composite designs and a 'radial' search, in order to reduce the search space and to estimate the experimental condition that minimizes the metric of interest (Average Browse Time in minutes). The value and location of the predicted optimum are:

Prev.Length	Match.Score	Prev.Type	Tile.Size	Avg. Browse Time
75	75	TT	0.2	10.0831

II. Introduction

Problem Statement

In this scenario, Netflix is experimenting with altering 4 features of their "Top Picks for ..." row on their desktop homepage with the goal of reducing the amount of time users spend choosing what to watch. By streamlining the browsing experience, the idea is to increase user engagement by reducing decision paralysis and fatigue. As users get frustrated by the number of choices, they may leave the site.

The goal is to discern the statistical significance and measurable impact of each of the four factors (**`Tile.Size`**, **`Match.Score`**, **`Prev.Length`**, **`Prev.Type`**), in order to discover the optimal combination of feature levels that minimizes browsing time, while being mindful of the cost associated with conducting experiments.

Experimental Methodology

The experiments can be divided into two steps: factor screening and response surface design. In order to avoid extraneous data collection, I first employed a 2^k factorial experiment to filter any insignificant factors from the search space. Although 2^k factorial experiments are useful for screening, they are limited in their response optimization capabilities due to their limited number of conditions.

To more rigorously search for the optimum within the search space, I conducted a central composite design (CCD) experiment. In 2 dimensions, this type of design requires data from 9 conditions, as opposed to the 4 conditions of a 2^k factorial design. These additional conditions allowed us to estimate the quadratic behavior of the response surface. From this, I was able to approximate the location of the minimum (stationary point) by setting the derivative of the estimated surface equal to zero.

The calculated stationary point fell outside of any previously observed conditions, and so I conducted a second CCD experiment centered around the first stationary point. With this, I was able to define the final search area, where I collected data on conditions 'radial' to the second stationary point.

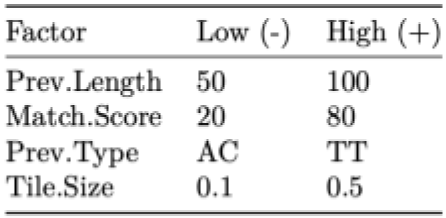
In this report, we will look more closely at the experimental journey, including the experimental designs, specific conditions, and analyses that led to the findings about the optimal conditions to minimize browsing time.

III. Experiments

Feature Screening

Problem/Design

The initial step was to reduce the search space by using a 2^4 factorial design, as it is a low cost way to determine which factors are significant in impacting average browsing time. The experimental conditions of the 2^4 factorial experiment are shown in *Table A*.

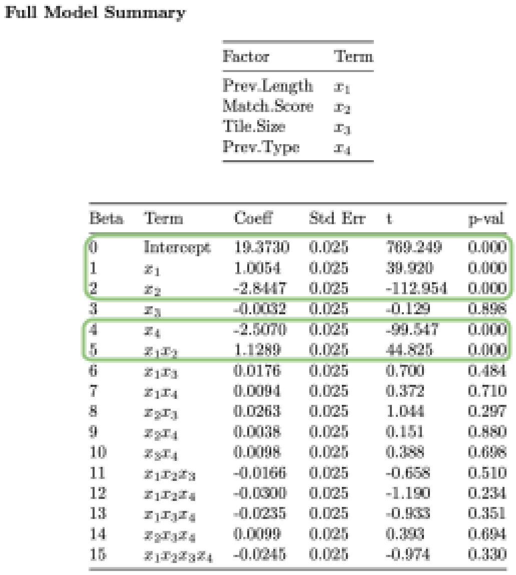


Factor	Low (-)	High (+)
Prev.Length	50	100
Match.Score	20	80
Prev.Type	AC	TT
Tile.Size	0.1	0.5

Table A

Analysis

From the results of the linear model as seen in *Table B*, the main effects of `Prev.Length`, `Match.Score`, and `Prev.Type` were significant at a 1% significance level as well as the interaction between `Prev.Length` and `Match.Score`.



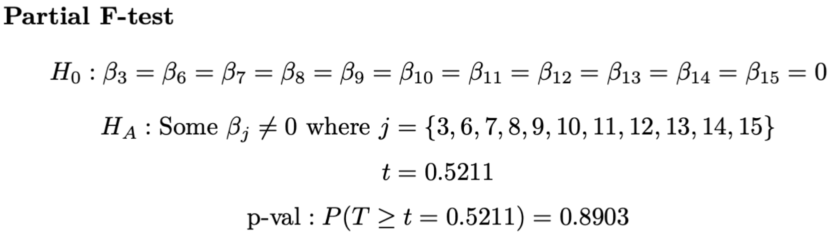
Full Model Summary

Factor	Term
Prev.Length	x_1
Match.Score	x_2
Tile.Size	x_3
Prev.Type	x_4

Beta	Term	Coeff	Std Err	t	p-val
0	Intercept	19.3730	0.025	769.249	0.000
1	x_1	1.0054	0.025	39.920	0.000
2	x_2	-2.8447	0.025	-112.954	0.000
3	x_3	-0.0032	0.025	-0.129	0.898
4	x_4	-2.5070	0.025	-99.547	0.000
5	x_1x_2	1.1289	0.025	44.825	0.000
6	x_1x_3	0.0176	0.025	0.700	0.484
7	x_1x_4	0.0094	0.025	0.372	0.710
8	x_2x_3	0.0263	0.025	1.044	0.297
9	x_2x_4	0.0038	0.025	0.151	0.880
10	x_3x_4	0.0098	0.025	0.388	0.698
11	$x_1x_2x_3$	-0.0166	0.025	-0.658	0.510
12	$x_1x_2x_4$	-0.0300	0.025	-1.190	0.234
13	$x_1x_3x_4$	-0.0235	0.025	-0.933	0.351
14	$x_2x_3x_4$	0.0099	0.025	0.393	0.694
15	$x_1x_2x_3x_4$	-0.0245	0.025	-0.974	0.330

Table B

To formally verify this, I ran a partial F-test comparing the full model to a reduced model that only included terms deemed significant by the full model. With a test statistic $t = 0.5211$ and p-value of 0.8903, we fail to reject the null (at a 1% significance level). Subsequently, I operated under the null hypothesis: $\beta_j = 0$ where $j \in \{3, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}$.



Partial F-test

$$H_0 : \beta_3 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = 0$$

$$H_A : \text{Some } \beta_j \neq 0 \text{ where } j = \{3, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}$$

$$t = 0.5211$$

$$\text{p-val} : P(T \geq t = 0.5211) = 0.8903$$

Because neither the main effect of **`Tile.Size`** nor any of its corresponding interactions were significant, it was determined that this was not a statistically significant factor. Consequently, **`Tile.Size`** was excluded from further explorations. Moreover, the main effect of **`Prev.Type`** was found to be statistically significant. However, the corresponding interactions were deemed insignificant, providing evidence of **`Prev.Type`**'s independence from the other significant factors (**`Prev.Length`**, **`Match.Score`**).

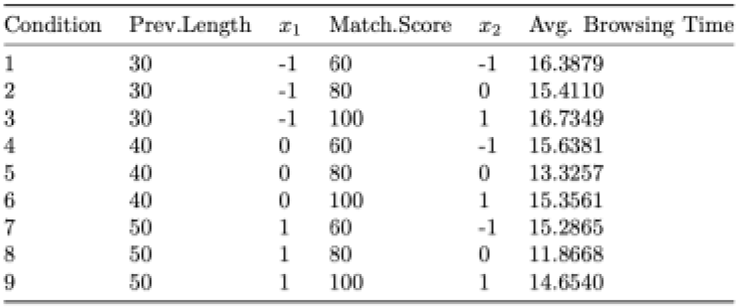
Due to its independence and that all levels of **`Prev.Type`** are considered in the model, we can use the results of the model summary to show that TT is the level of **`Prev.Type`** that minimizes average browsing time. Since $\beta_4 =- 2.5070$, the bounds of the 95% confidence interval are both negative, and TT was coded as +1, we can conclude that the expected value of average browsing time changes by $2\beta_4 =- 5.014$ when changing **`Prev.Type`** from AC to TT, holding all else equal.

Using a 2^4 factorial design, we've successfully screened the factors and reduced the search space to two factors (**`Prev.Length`** and **`Match.Score`**). However, due to the limited number of levels, this type of experimental design is not suited for response optimization. To locate the optimum more systematically, I employed central composite design experiments.

Response Surface Designs

Design (CCD 1)

With the goal of locating the optimum, we impose quadratic behavior on the model to take advantage of the concavity of the estimated response



Condition	Prev.Length	x_1	Match.Score	x_2	Avg. Browsing Time
1	30	-1	60	-1	16.3879
2	30	-1	80	0	15.4110
3	30	-1	100	1	16.7349
4	40	0	60	-1	15.6381
5	40	0	80	0	13.3257
6	40	0	100	1	15.3561
7	50	1	60	-1	15.2865
8	50	1	80	0	11.8668
9	50	1	100	1	14.6540

Table C

curve. To collect sufficient data to estimate quadratic behavior, we need $(K' + 1)(K' + 2)/2$ experimental conditions (6 in this case). Since I followed a central composite design, the 9 conditions were more than enough. The experimental conditions of the central composite design are shown in *Figure A* and *Table C*. I chose these conditions instead of centering at the estimated optimum because 2^k factorial experiment designs have a relatively poor capacity for response optimization and I wanted to explore the extreme values of both factors. For the axial conditions, I used $a = 1$ to minimize data collection, as this includes data collected previously[1].

Analysis (CCD 1)

From the quadratic model of the response surface, I found the stationary point to be $[2.0366,\ 0.1385]^T$ as shown in the contour plot of *Figure B*. This corresponds to $[60.3655,\ 82]^T$ in natural units. Because the stationary point fell outside of previously observed experimental conditions, I suspected that the estimate varied from the true optimum. Therefore, it was decided another central composite design experiment oriented around this stationary point would provide a more robust estimation.

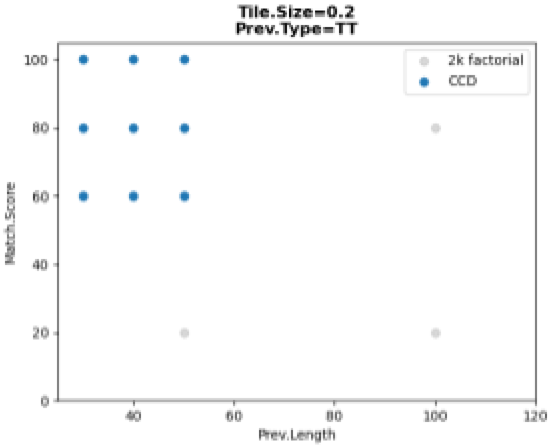


Figure A

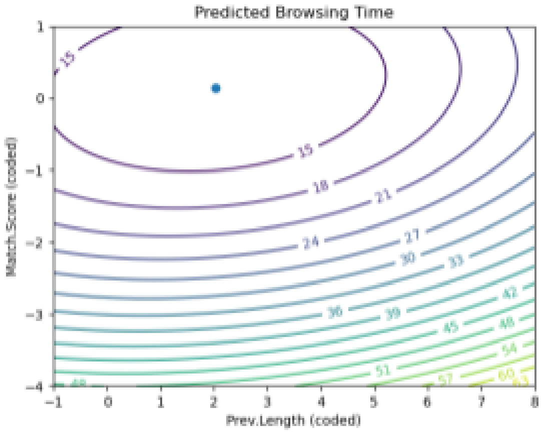
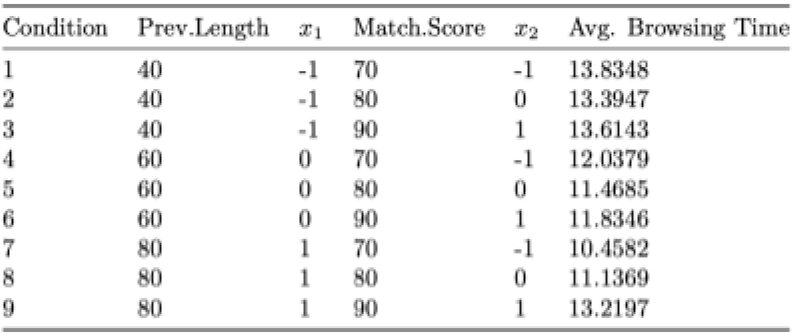


Figure B

Design (CCD 2)



Condition	Prev.Length	x_1	Match.Score	x_2	Avg. Browsing Time
1	40	-1	70	-1	13.8348
2	40	-1	80	0	13.3947
3	40	-1	90	1	13.6143
4	60	0	70	-1	12.0379
5	60	0	80	0	11.4685
6	60	0	90	1	11.8346
7	80	1	70	-1	10.4582
8	80	1	80	0	11.1369
9	80	1	90	1	13.2197

Table D

In the second CCD experiment, I chose to center the experimental conditions around $(60,\ 80)$ because it is a close and realistic set of values for each factor under the business restrictions. The experimental conditions are also

[1] However, we failed to account for the default values of the other factors when collecting the new data, so we ended up having to recollect the previous condition, thus not saving resources.

shown in *Table D* and *Figure C*. I chose these conditions to gain information within the lowest contour of the contour plot of the previous CCD experiment (*Figure B*). Additionally, these specific values allow us to reuse a condition from the first CCD.

Analysis (CCD 2)

From the new model of the response surface, I found that the new stationary point lies at $[1.1744, -1.2654]^T$ as shown in the contour plot in *Figure D*. This corresponds to $[83.4886, 67.3458]^T$ in natural units.

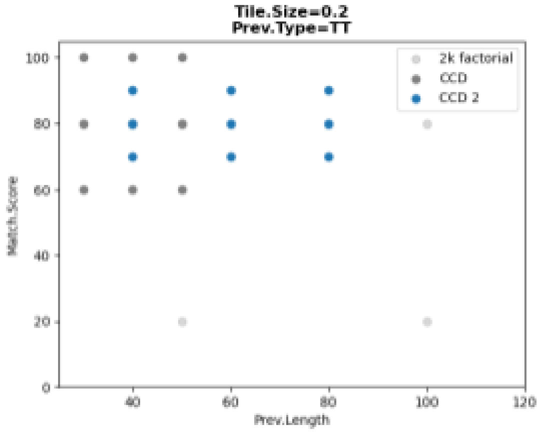


Figure C

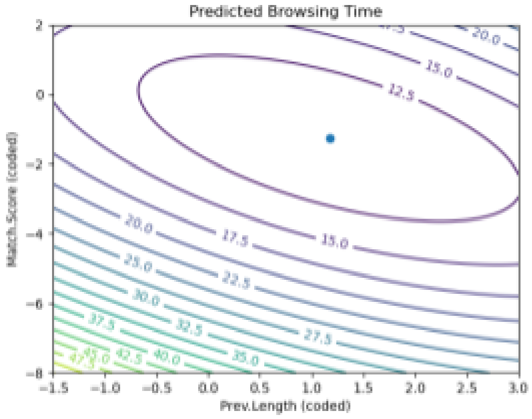
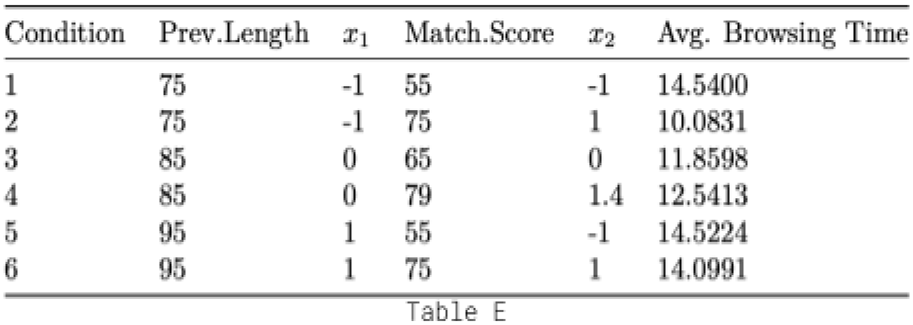


Figure D

'Radial' Search

In order to improve the chances of locating the minimum, I exhausted the rest of the experimental units (40 total) and collected data in the fashion of a 2^2 factorial design with one axial condition, centered around $(65, 85)$. Despite the use of a design similar to 2^k factorial, the goal at this stage was to systematically search the area around the previously estimated minimum, not to screen out insignificant factors. For the axial condition, I used $a = 1.4$, projecting upwards from the center. The conditions for which data were collected are shown in *Table E*. I chose these intervals to look into multiples of 5, since the search had previously been limited to multiples of 10.



Condition	Prev.Length	x_1	Match.Score	x_2	Avg. Browsing Time
1	75	-1	55	-1	14.5400
2	75	-1	75	1	10.0831
3	85	0	65	0	11.8598
4	85	0	79	1.4	12.5413
5	95	1	55	-1	14.5224
6	95	1	75	1	14.0991

Table E

IV. Conclusion

The goal for this series of experiments was to discover which of the four factors of the Netflix homepage (**`Tile.Size`**, **`Match.Score`**, **`Prev.Length`**, **`Prev.Type`**) are statistically significant in contributing to average browsing time and to measure the impact of specific changes in these factors.

In the screening step, using a 2^k factorial design, I was able to identify **`Prev.Type`** as a statistically significant, but independent, factor and **`Tile.Size`** as an insignificant factor. This allowed us to immediately determine the optimal level of **`Prev.Type`** and eliminate **`Tile.Size`** from consideration. Being able to draw these conclusions helped us to reduce the dimensionality for the subsequent response surface design experiments to just two factors: **`Prev.Length`** and **`Match.Score`**

In the response optimization step, using two central composite design experiments, I was able to calculate stationary points from quadratic response surfaces to close in on the values of **`Prev.Length`** and **`Match.Score`** that minimize average browsing time

Finally, I expanded the search around the resulting stationary point to find the proposed experimental condition (**`Prev.Length`**: 75, **`Match.Score`**:75, **`Prev.Type`**: TT, **`Tile.Size`**: 0.2)that minimizes browsing time at 10.0831 minutes.

Like all optimization searches, the limitations relate to the search space. Since there were limits imposed on conducting experiments, the search is very dependent on the specific levels of each factor I choose to explore. Given more resources, I could dive deeper to either confirm the estimate or possibly discover a global minimum elsewhere.

If I were to continue to optimize the homepage, the next step would be to consider estimating the response surface with all of the data points collected, and deriving a possible optimum from this.