MSDS 629 Final Project

Maricela Abarca, Belinda Ong, Lance Santerre, Ho Nam Tong

Component #1: Executive Summary

When faced with too many viewing options, Netflix users may become overwhelmed and ultimately not use our product. A series of experiments and regression models were used to find the optimal combination of factors to minimize Netflix's homepage browsing time, and therefore reduce the risk of users becoming overwhelmed. The optimum browsing time was found to be a match score of 75, preview length of 75 seconds, and preview type of Teaser Trailer ('TT'), resulting in a predicted average browsing time of 9.96 minutes, with a 95% confidence interval of 7.86 to 12.06 minutes. Tile size does not significantly influence average browsing time.

Component #2: Introduction

By minimizing browsing time, the risk of losing user interest due to excessive viewing options is reduced. Four factors that may influence time spent browsing the homepage are tile size, match score, preview length, and preview type for offerings in the “Top Pick For…” row of the homepage. Through multivariate experiments, an optimal combination of these factors can be found that minimizes average browsing time and elevates user experience.

The first step to finding optimal factor levels is to screen factors for significant impact on browsing time. A 2^k + Center Point factorial design was implemented for this purpose, with the addition of center point condition to simultaneously test for response curvature. The levels used can be found in Table 1. The results showed significance for match score, preview length, and preview type, but not for tile size. A test for curvature showed that these initial screening levels were in fact in an area of curvature.

Based on results from the first experiment, a central composite design was implemented to find the optimal combination of significant factors. Second order regression models were used to locate theoretical minima for two separate surfaces for the categorical levels of preview type. The teaser/trailer surface contained the smallest average browsing time, so optimal levels of match score and preview length were calculated from this model. These were translated to practical levels of match score and preview length. A confirmation experiment was conducted at these levels to collect data around this point.

As a final check, we calculated the actual average browsing time for each condition collected across all experiments. When we noted that the actual average browsing time was lower for a condition different from the model identified minimum, we conducted a one-sided t-test to confirm that the difference is significant. Based on that, we finalized our optimized condition and calculated the 95% confidence interval for average browsing time.

Component #3: The Experiments

3.1 Experiment 1: 2^k + Center Point factorial design

The objective of this first experiment was to screen for significant factors and verify if the levels used for screening were in an area of curvature. Screening tests often do not occur over a design space that contains an optimum but in order to confidently rule out the screening levels, a 2^k + center point design was implemented. Including a center point condition in the 2^k factorial design allowed for a test of curvature around the screening values. The screening values can be found in their natural and coded units in Table 1. Eighteen total conditions were tested in this first experiment with 100 users randomly assigned to each condition .

Condition*	Tile Size	Coded Size	Match Score	Coded Score	Preview Length	Coded Length
1	0.1	−1	50	−1	30	−1
2	0.1	−1	50	−1	120	1
3	0.1	−1	100	1	30	−1
4	0.1	−1	100	1	120	1
5	0.5	1	50	−1	30	−1
6	0.5	1	50	−1	120	1
7	0.5	1	100	1	30	−1
8	0.5	1	100	1	120	1
9	0.3	0	75	0	75	0

Table 1: Experimental conditions for the factor screening and curvature test.
*There were 18 total conditions, as these 9 conditions were used for each of the two levels of Preview Type: Actual Content and Teaser Trailer.

Tile size and preview length were screened at their minimum and maximum values within their regions of operability, which maximized the chance of finding significant effects if they existed. Instead of using the minimum value of 0% for match score, we decided to use a 50% score based on the assumption that scores between 0% and 49% would likely not be put into practice. Testing with these values could also potentially negatively impact users' attitudes toward recommendations, and therefore Netflix products.

After fitting a full Ordinary Least Squares (OLS) model and applying a 5% significance threshold, we found that tile size and its interactions did not have a statistically significant impact. Match score, preview type, and preview length were significant (p-value 0). The interaction between match score and preview length was also significant (p-value 0). The results of fitting a second order model with these significant effects plus the one significant interaction showed significant coefficients for the quadratic terms (both p-values 0), indicating curvature in the surface and that a possible optimum is nearby.

Based on these findings, subsequent experiments excluded tile size and all interaction terms except for the one between match score and preview length. Screening levels were also subsequently used to explore optimal factor levels of match score and preview length.

3.2 Experiment 2: Central Composite Design

Because the first experiment confirmed that the initial screening levels are in an area of significant curvature, the data collected from the first experiment were augmented with data from axial conditions that would allow for a face-centered Central Composite Design with a = 1. The new conditions in their natural and coded values are in Table 2. One hundred users were randomly assigned to each of the eight axial conditions.

Condition*	Match Score	Coded Score	Preview Length	Coded Length
1	75	0	30	−1
2	75	0	120	1
3	50	−1	75	0
4	100	1	75	0

Table 2: Additional experimental conditions for the Central Composite Design experiment.
*There were 8 total conditions, as these 4 conditions were used for each of the two levels of Preview Type: Actual Content and Teaser Trailer.

The results of the experiment from these eight new conditions were combined with the results from the first experiment to create a dataset that would enable the fitting of a second order model that would locate the minimum browsing time efficiently with a larger dataset of conditions. The data were separated in two groups for the categorical variable preview type: Teaser/Trailer (TT) and Actual Content (AC). Because the levels for preview type are categorical, coefficients cannot be estimated in the same way as numerical factors with one second order regression model. Instead, two second order models for TT and AC were fitted separately and the results were compared to identify the minimum between the two models.

For each model, contour plots were created to visualize and estimate the optimum levels and value of the minimum average browsing time at the stationary point (Figure 1). The stationary point was then calculated by constructing matrices and applying the optimum expected response formula. After translating these findings into real-world units, they were run through the models to estimate average browse time for both the TT and AC surfaces.

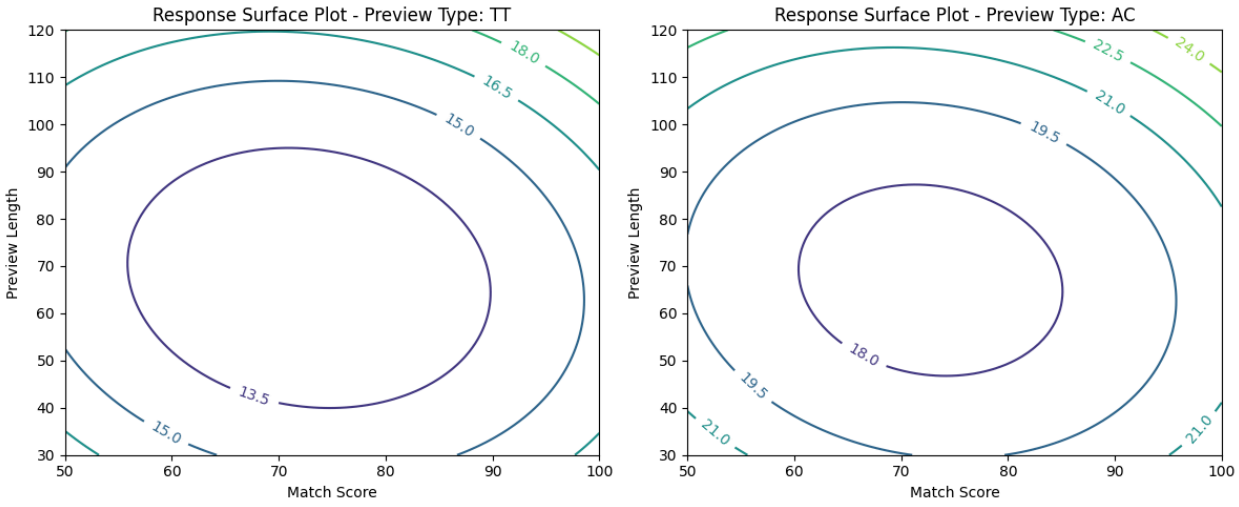

Figure 1. Contour plots in natural units for the two preview type levels, indicating that TT has a smaller minimum.

The theoretical optimal average browse time was calculated to be 12.34 minutes for the TT model and 17.39 minutes for the AC model. This illustrates the efficiency of the TT model in reducing browsing time. Specifically, this finding suggests that using a teaser trailer, as opposed to actual content, effectively lowers average browse time.

For the TT model, the stationary point occurred at a match score of 72.8% and a preview length of 67.5 seconds. A prediction function was employed to explore variations in these input values. This function was used to assess the impact of small adjustments to these input values, a necessary step due to practical constraints (i.e., match score must be an integer and preview length can only change in increments of 5 seconds). This exploration revealed a practical minimum average browsing time of 12.35 minutes at a match score of 73% and a preview time of 65 seconds.

Therefore, the model optimum occurs at a preview type TT with a match score 73% and a preview length of 65 seconds.

3.3 Confirmation Experiment and Confidence Interval

We ran a confirmation experiment with the identified model optimum condition from section 3.2 to collect data for a confidence interval and actual mean browsing time. The actual mean under match score 73% and preview time 65 seconds was 10.76 minutes with a 95% confidence interval of [8.80, 12.68], which is even smaller than the model predicted mean of 12.35 minutes.

These results made us confident in our findings of a minimum location, but as a final check, we calculated the actual average browsing time for each condition already collected across all experiments to make sure we hadn't collected data that suggested an even smaller minimum average browsing time. A new minimum average browsing time was in fact identified among all the experimental conditions at 9.96 minutes (Actual Minimum in Figure 2) when match score is 75%, preview length is 75 seconds, and preview type is TT. This is the actual minimum from our collected data from the first round of experimental conditions described in section 3.1.

This mean is also smaller than our model optimum condition, which resulted in an average actual browsing time of 10.76 minutes (Model Minimum in Figure 2). Therefore, we conducted one final experiment at the next logical point along the same line at match score 76%, preview length 80 seconds, and preview type TT to confirm that the average browsing time is larger at that point. This would imply that we've gone beyond the area of the minimum. This final testing point yielded an average browsing time of 10.49 minutes (Testing Point in Figure 2). These results support that the actual minimum average browsing time of 9.96 minutes from the experimental conditions at match score 75% and preview length 75 seconds is the true minimum location and we should use these levels as the optimal factor levels instead of the model identified optimal levels.

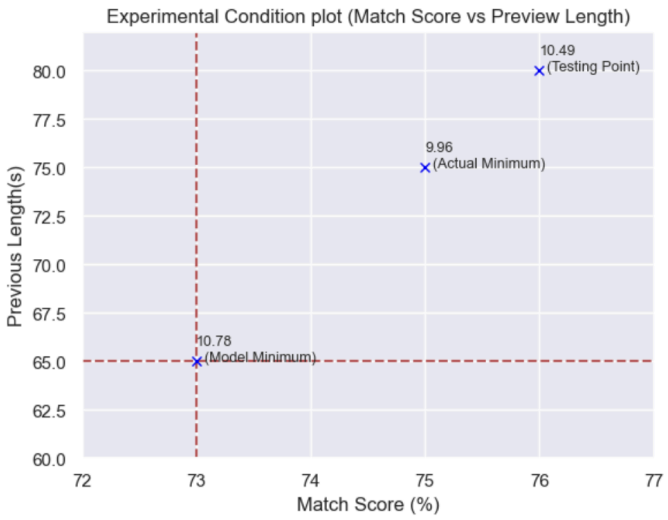

Figure 2. Experimental condition plot describing the minimum browse time under different conditions.

Finally, we conducted a one-sided t-test to confirm that the Actual Minimum is smaller than the Model Minimum (Figure 2). Based on that test (p-value 8.240 x 10^{-8} , 5% significance level), we concluded that the Actual Minimum has the shorter average browsing time. We also calculated the 95% confidence interval for the Actual Minimum and concluded that the average browsing time lies between 7.86 minutes to 12.06 minutes. Table 3 summarizes the three combinations of factor levels identified throughout our process to be potential locations of average minimum browsing time.

Factor Levels (all held at TT)	**Average Browse Time (minutes)**	**Confidence Interval (average browse time)**
Match Score 75, Preview Length 75*	9.96	[7.86, 12.06]
Match Score 76, Preview Length 80	10.49	[8.39, 12.59]
Match Score 73, Preview Length 65	10.76	[8.80, 12.68]

Table 3: Candidate factor levels for minimum average browsing time.*Levels selected as optimum condition.

Component #4: Conclusion

Through a sequential set of experiments and analysis, our study examined four factors of the "Top Picks For…" row of the Netflix homepage in order to minimize average browsing time. Our results suggest that this row of the homepage should employ a match score of 75%, preview length of 75 seconds, and teaser/trailer content in order to keep average browsing time at 9.96 minutes with a 95% confidence interval of 7.86 minutes to 12.06 minutes. Tile size was shown to not significantly affect browse time, so can be selected to aesthetically complement the homepage within the range of a 0.1 to 0.5 ratio of tile height to screen height.

Despite our comprehensive approach, we acknowledge that limitations might affect our findings. One such limitation is that there are likely uncontrolled sources of variation in our designs that affect average browsing time, such as time of day, day of the week, user demographics, etc. Additionally, we only experimented with the "Top Picks For…" row of the homepage, but other aspects of the homepage likely affect how long users spend browsing for content. We also only experimented with four factors in this row. Future experiments could explore other sources of variation and factors not covered by our study, such as the actual algorithm that computes match score and if different variations of that algorithm optimize browsing times.

Other limitation include: (1) Only 28 experimental runs were performed and more data points near the identified optimum may further improve precision; (2) practical constraints such as match scores can only be integers and preview lengths in increments of 5 seconds which may prevent finding the true unconstrained optimum; (3) quadratic effects were modeled, but higher order polynomial terms were not estimated so the true response surface may not follow a quadratic model; (4) the long term generalization of our findings was not estimated, so these optimal factor levels may not endure over long periods of time.