Executive Summary

Netflix is a streaming service that allows users to purchase subscriptions to a large variety of TV shows, movies, documentaries and more. With this vast library of viewing options, users can become overwhelmed, causing decision-paralysis. To remedy this, three factors were explored to determine the optimal combination that minimized the average time spent browsing the homepage. These three factors were match score, a prediction of how much a user will enjoy watching the show or movie based on their viewing history; preview length, the duration of a show or movie's preview; and tile size, the ratio of a tile's height to the overall screen height.

Response surface methodology was used to conduct a series of sequential experiments to transfer knowledge gained from one experiment to the next. Through factor screening, it was determined that match score and preview length were factors that significantly affected the average browsing time; however, tile size and its corresponding interactions did not. This was determined by fitting a full factorial model with the three factors to conduct a hypothesis test on their significance; these respective p-values were compared against a 1% significance level. Hence, for future experiments, only match score and preview length were considered.

Next, to determine the vicinity of the optimal combination of these values, the method of steepest descent was conducted in two rounds. The first one began at the original center point of 110 seconds for preview length and a 90% match score. The optimum was centered at 85 seconds for preview length and 64% for match score after six steps; this resulted in an average browsing time of 11.85 minutes. The optimum was identified by calculating the gradient and then comparing the average browsing time at each step of the descent to identify when it was no longer decreasing. The goal of the next round was to recompute the gradient to identify a more precise region of the optimum beginning at the optimum from the first round; this resulted in a gradient in another direction. The final region of the optimum was centered at 80 seconds for preview length and 73% for match score after two steps resulting in an average browsing time of 10.38 minutes. Lastly, to ensure we were in the vicinity of the optimum, a 2^2-factorial experiment with center point was conducted to test for curvature by fitting a linear regression model. This was done using a hypothesis test which provided strong evidence to reject the null hypothesis that there was not curvature; hence, we had achieved curvature and identified our region of the optimum for response optimization.

Lastly, through response optimization, we were able to determine the optimum preview length and match score. This was done by conducting a central composite design experiment using the same boundaries as round 2 of the steepest descent. This involved using a 2^2-factorial experiment with center point and axial conditions. The axial conditions were determined by setting $\alpha = \sqrt{2}$ since the region of the optimum was not close to any of the factor's region of operability boundaries. Thus, we used a rotatable design for our response optimization. By fitting this second-order surface model, the stationary point (optimum) was identified in natural units to be 75.5936 seconds for preview length and 74.42077% for match score. Hence, a more practical and feasible conclusion would be to set 75 seconds for the preview length and 74% for the match score to minimize average browsing time. This yielded an estimated average browsing time of 10.01 minutes and a 95% confidence interval of (9.85, 10.16) minutes.

Introduction

Netflix is streaming service that allows users to purchase varying subscriptions to watch TV shows, movies, documentaries and more on their device. This large variety of shows/movies are organized into tiles on a grid system on the homepage. The different rows correspond with recommendations, countries, or genres to categorize the homepage. Netflix has revolutionized streaming services and is known for their recommendation systems which allow them to provide catered top picks for each of their users based on their history – as users watch more, Netflix is able to provide more accurate recommendations.

With this vast variety of TV shows and movies, users can face the problem of decision-paralysis when deciding on what to watch. Ultimately, the choice-overload and difficulty with deciding can result in users losing interest and logging off the platform. Hence, to address this problem, Netflix would like to further investigate an optimal solution to alleviate this fatigue from browsing the choices on the homepage.

Netflix tries to overcome this issue through methods that help facilitate quicker decisions. This can include, but is not limited to, providing recommendations based on the user's previous watching history, providing previews of the shows/movies, and/or increasing the tile sizes so the previews are larger. Previews are teasers/trailers that are enlarged and automatically played when hovering over the tile. For manipulating the tile size, the aspect ratio is fixed so the size can be altered, but the shape remains the same. The table below summarizes the factor's description, their region of operability and the levels of their experiment.

Table 1: Factor descriptions and relevant information

Name	**Description**	**Region of Operability**	**Levels**
Preview Length (A)	The duration (in seconds) of a show or movie's preview	[30,120]	Low: 100 High: 120
Match Score (B)	A prediction of how much a user will enjoy the show or movie, based on their viewing history	[0,100]	Low: 80 High: 100
Tile Size (C)	The ratio of the tile's height to the overall screen height.	[0.1, 0.5]	Low: 0.1 High: 0.3

For the purposes of this experiment, the three factors listed above will be investigated with the "Top Picks For…" row of the homepage. This is the row that contains personalized recommendations based on a user's past viewing history.

We are interested in measuring the time spent browsing the homepage browsing time and how that can be minimized based on these factors. The goal is to optimize these factors to peak a users' interests in a show and allow them to make decisions with ease; thereby, resulting in less time spent browsing the homepage. Hence, our metric of interest (MOI) is the average browsing time, and our corresponding response variable is the browsing time, in minutes. Each

experimental unit is a user that has been assigned to one of the conditions for their Netflix homepage.

This experiment will be conducted using response surface methodology. This is a set of sequential experimentation that utilizes information gained in previous experiments to inform decisions made in future experiments. First, we will conduct factor screening to determine which, if any, of these factors significantly affect the average browsing time. The information gained from this first phase will be used to narrow down the factors such that resources for testing and quantifying results are only allocated to significant factors. Once these factors are decided, the next phases are for response optimization. This will involve the method of steepest descent and response surface designs to determine the optimal combination of our significant factors to minimize average browsing time. The information from the method of steepest descent will provide us the vicinity of the optimum. This will be extremely useful for the response surface designs as these approximations would be poor across the entire region of operability; however, within the localized region of the experiment, these functions will approximate our values well. After these three phases of experimentation, we will arrive at a recommendation for Netflix regarding how these factors should be set to minimize the average browsing time.

Factor Screening

The objective of this phase is to determine which factors significantly affect browsing time. This was designed using a 2^3-factorial experiment; the conditions were all combinations of the high/low levels of the three factors. This design was preferred over a fractional factorial experiment since with three factors, aliasing would involve two-factor interactions (i.e., A=BC). In each case, we are associating the main effect of one factor with a two-factor interaction, which could prove to be significant. To avoid these complications, the full factorial design was run.

A total of 100 users were assigned to each of the 8 conditions and their browsing time data was collected. A full linear regression model was fit since the average browsing time was our metric of interest. This involves the main effects of the factors, two-factor interactions, and a three-factor interaction. The p-values associated with each of the effects and interactions are summarized in Table 2. These p-values are calculated for the hypothesis test to determine if the associated β for each effect is 0. The null hypothesis is that the relevant β is 0 whereas the alternative hypothesis is that the β is non-zero. If the p-value is less than our 1% threshold, we can reject our null hypothesis; thus, showing that the β is non-zero and our main/interaction effect is significant. A stricter 1% significance level was chosen for comparison to ensure higher significance rather than using a broader 5% significance level.

Table 2: Hypothesis test results for each effect

Factor	**p-value**		**Interaction**	**p-value**		**Interaction**	**p-value**
A	< 2e-16		A:B	< 2e-16		A:B:C	0.9715
B	< 2e-16		A:C	0.1236			
C	0.0456		B:C	0.3616			

From the summary in Table 2, the active factors and two-factor interactions are A, B and A:B. Factor C and it's respective two and three-factor interactions are not significant as their p-values are above 1%. In the context of this problem, if we switch from the high to low factors for preview length or match score, then the average browsing time is significantly affected. The main effects of the two active factors are summarized in Table 3.

Table 3: Main Effects of Preview Length and Match Score

Factor	**β**	**Main Effect**	**Interpretation**
A (Preview Length)	3.06	6.12	When we switch from a preview length of 100 seconds to 120 seconds, we expect the average browsing time to *increase*.
B (Match Score)	3.11	6.22	When we switch from a match score of 80 to 100, we expect the average browsing time to *increase*.

For the active factors and two-factor interactions, the main effect and interaction effect plots are below.

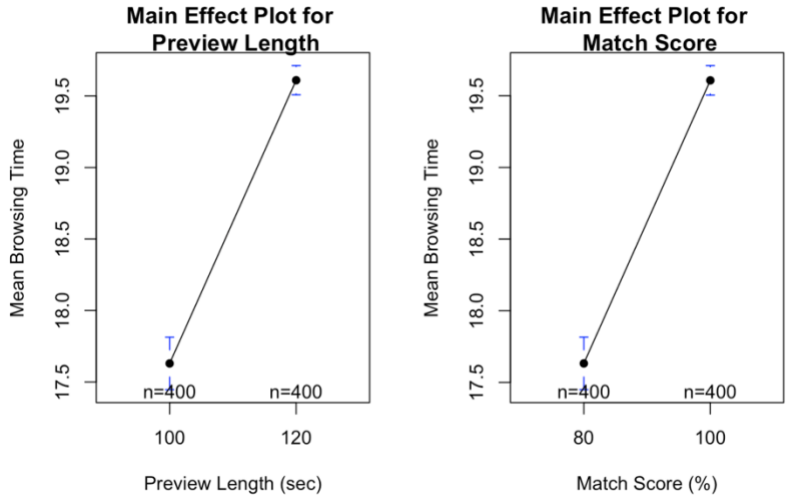

Figure 1: Main Effect Plot for Preview Length and Match Score

From the plots above, we observe that the results match the interpretation of Table 3. It should be noted that the preview length time is not included in the browsing time measurement. The higher preview length could show more content resulting in information overload and a higher browsing time. This higher match score might result in inaccurate recommendations when catered to each user as a 100% match is unlikely with a prediction model. An alternative theory is that it may recommend content that is too similar to previously watched shows/movies. Both plots have a relatively similar influence on the average browsing time as we switch between levels.

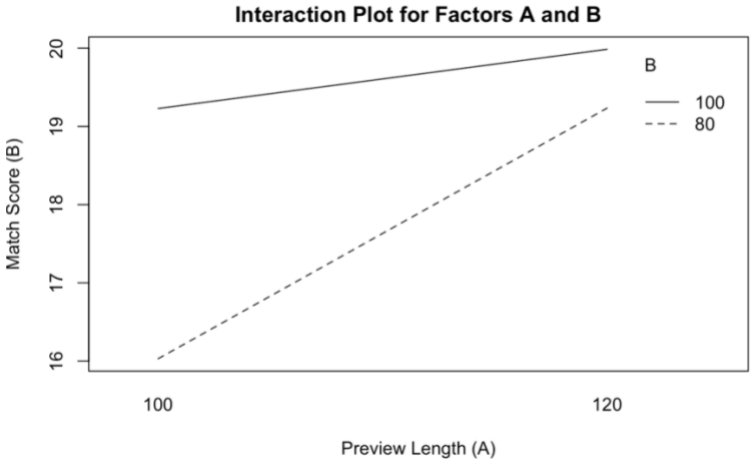

Figure 2: Interaction Plot for Preview Length and Match Score

Since the lines in Figure 2 are not parallel, we can conclude that the interaction is significant. The optimal combination that minimizes average browsing time from this experiment is a preview length of 100 seconds and an 80% match score.

Therefore, from our factor screening process, the tile size was concluded to be insignificant and will be set to a default value of 0.2 for the subsequent phases of experimentation. Match score, preview length and their interaction were proven to significantly affect the average browsing time and will be investigated further in follow-up experiments to determine the optimal values of these two factors to minimize the MOI.

Method of Steepest Descent

The aim of this method is to determine the vicinity of the optimum for our significant factors. This will be used in the next phase to provide a localized region to explore. The path of steepest descent is identified using a 2^2-factorial experiment with a center point condition to estimate our first-order response surface; we descend the surface to minimize our MOI. The center point in natural units is 110 seconds and a 90% match score. Our step size will depend on preview length since this is a more difficult factor to manipulate; the factor step size is set to 0.5 (coded units) to ensure steps of 5 seconds. Then to get our overall step size for our gradient, we will divide that step size by the respective absolute value of β-coefficient. Data is collected for every step to determine when average browsing time is minimized. At the location of the best MOI value, a 2^2 -factorial experiment will be conducted to test for curvature. Initial data was collected for conditions with preview length (100, 100, 120 seconds) and match score (80, 90, 100%).

It is important to note that match score must be an integer and preview length can only be altered in increments of 5 seconds. Hence, at any stage of our steepest descent, these two factors will be rounded accordingly. The following figure displays the steps and direction for the first round of the method of steepest descent:

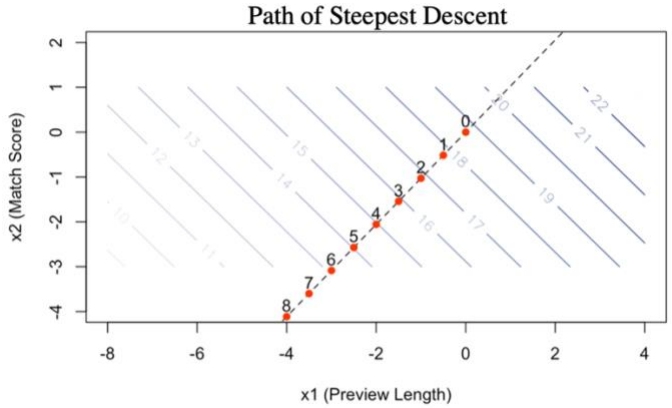

Figure 3: Path of Steepest Descent (Round 1)

Hence, we can see that the path is descending towards the bottom-left corner. For each of these combinations of the preview length and match score, 100 users were assigned to the condition and their browsing time data was collected. The average browsing time of each condition is summarized in Table 4 and Figure 4.

Table 4: Average Browsing Time (Round 1)

Step	Preview Length (seconds)	Match Score (%)	Avg. Browsing Time (minutes)
0	110	90	19.23
1	105	85	18.12
2	100	80	16.24
3	95	75	14.18
4	90	69	12.15
5	85	64	11.85
6	80	59	13.11

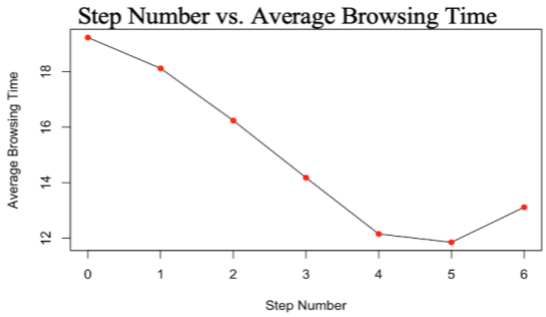

Figure 4: Step Number vs. Average Browsing Time

From Table 4 and Figure 4, we observe that the average browsing time is minimized at Step 5 whereby our preview length was 85 seconds, and our match score was 64%, resulting in an average browsing time of 11.85 minutes. For this round of steepest descent, step 5 is our stopping point; if we were to carry further with another step, our average browsing time would increase. To fine-tune this optimum further, we recalculate our gradient with levels at step 4, 5, and 6 to determine if a descent in another direction would bring us closer to the optimum. This will require more resources to pull data at each new step; however, if another direction would result in getting a closer estimate, this is worth the trade-off. These steps were used as the levels to ensure a 5 second increment between the preview length conditions and to reuse data from round 1 for the gradient calculation. Table 5 displays the steps taken with this new gradient.

Table 5: Average Browsing time (Round 2)

Step	**Preview Length (seconds)**	**Match Score (%)**	**Avg. Browsing Time (minutes)**
0	85	64	11.85
1	80	73	10.38
2	75	82	10.91

In this case, our average browsing time was minimized in the first step. As we can see, the direction of our gradient changed as our match score began to increase while preview length decreased. Hence, our optimum is in the vicinity of 80 seconds for preview length and 73% match score resulting in an average browsing time of 10.38 minutes. These steps will be used to provide the localized region in the response optimization.

Next, we conducted a 2^2-factorial experiment with a center point to test for curvature by fitting a linear regression model. This data includes the following levels (high, center, low): preview length at 75, 80, 85 seconds and match score at 64, 73, and 82%. These levels were used to maintain the 5 second increments for preview length and use the corresponding match score when testing for curvature. This also allowed us to save on resources by reusing the data from the second round of steepest descent. This curvature test is a hypothesis test with H_0: $\beta_{PQ} = 0$ and H_A: $\beta_{PQ} \neq 0$. In this case, β_{PQ} is the pure quadratic effect which determines if our response values in our factorial conditions are similar to those in the center point condition. It is important to note that this test is assuming that our estimated quadratic β-values have the same sign for preview length and match score. Our associated p-value for this test is 3.912457×10^{31}. Thus, we can reject our null hypothesis which means curvature is achieved. Therefore, our region of the optimum centered around 80 seconds and match score of 73%.

For the next phase of response optimization, the boundaries for the second round of steepest descent will be used to mark the region of the optimum.

Response Optimization

The objective of this phase is to identify the location of the optimum for preview length and match score. We begin with using our region of the optimum that was discovered in the method of steepest descent as it provides us with a small, localized region that we have proven to contain the true optimum. These optimal values will be determined by conducting a central composite design (CCD) and then fitting a second order response surface model to determine the values that minimize average browsing time. This CCD requires running a full 2^2-factorial design with center point and axial conditions and then fitting a linear regression model.

The low and high levels of the CCD use the same boundaries as the second round of steepest descent to maintain the 5 second increments for preview length. These axial conditions require the identification of an adequate value of α. Since our region of the optimum is not near a corner of the region of operability for any of the factors, we do not use $\alpha = 1$. Instead, we set $\alpha = \sqrt{2}$ since we have 2 factors to ensure that the estimate of our response surface at every condition is equally precise; this provides a rotatable design. With regards to collecting data, the factorial experiment and center point data are reused from phase 2 to save resources; the axial conditions required new data to be collected. The axial conditions are at $\alpha = \pm \sqrt{2}$ while setting the other factor to the 0 in coded units. Hence, with these conditions determined, we can fit our full second order response surface model and identify the stationary point – our optimum. In the natural units, we have the following average browsing time for each condition summarized in Table 6.

Table 6: Average browsing time for 2^2-factorial design with center point and axial conditions

Condition	Preview Length (seconds)	Match Score (%)	Avg. Browsing Time (minutes)
1	75	64	10.99
2	75	82	11.11
3	85	64	10.23
4	85	82	12.74
5	80	73	11.85
6	87	73	11.98
7	73	73	12.08
8	80	86	12.08
9	80	60	11.93

Then, with this data, the second-order response linear regression model is fit. The contour plot of the fitted response surface is shown in Figure 5 in coded units and Figure 6 in natural units.

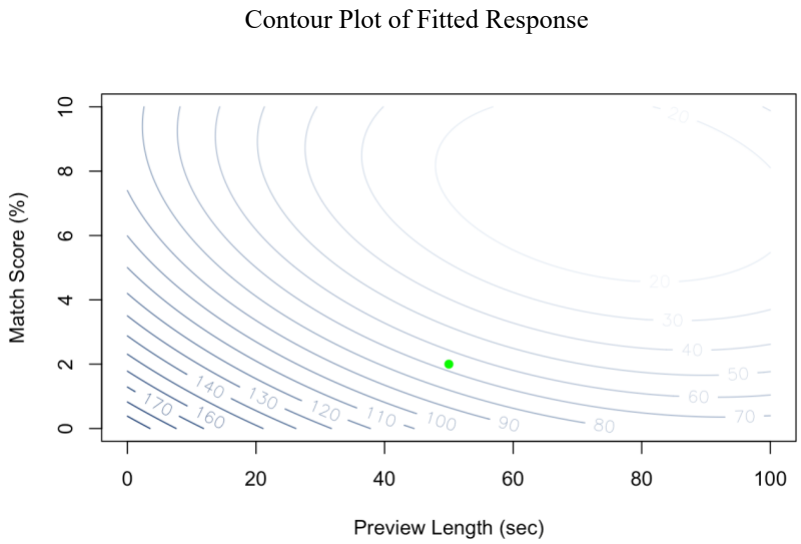

Figure 5: Contour plot of fitted second-order response surface (coded units)

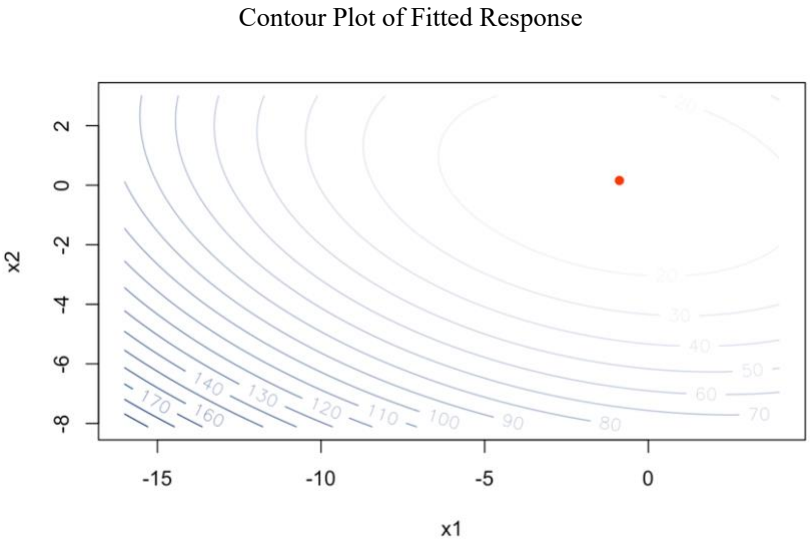

Figure 6: Contour plot of fitted second-order response surface (natural units)

The stationary point from this second order model is located (in coded units) at (-0.88, 0.16). We have identified our optimal location in natural units to be 75.5936 seconds for preview length and 74.42077% for match score. Hence, rounding these to the 5 second increments for preview length and integer value for match score provides us with an optimum of 75 seconds and 74%. This is a feasible combination for Netflix to set for preview length and match score to minimize the average browsing time.

This combination yields an estimated average browsing time of 10.01 minutes and a 95% confidence interval of (9.85, 10.16) minutes.