Executive Summary

The purpose of this experiment was to identify which factors significantly influence Netflix browsing time and determine which combination of these factors' levels will minimize the average browsing time. According to the Factor Screening Design, match score and preview length were found to be significantly influential. The Method of Steepest Decent was used to move from the initial region of experimentation with a center point (Preview Length, Match Score) = (110,90) towards the vicinity of the optimum and to locate optimal settings of these two factors. The contour plot for the estimated first order response surface was produced to visualize the path of steepest descent. According to the Method of Steepest Decent procedure, the optimal preview length was estimated to be in the vicinity of 90 sec and the optimal match score in the vicinity of 67%. Next, a test of curvature was performed in the region with center point (Preview Length, Match Score) = (90, 67) to determine whether the vicinity of the optimum was reached. A full second order response surface model fas fit to identify the factor levels that minimize expected browsing time. Central Composite Design was used to investigate 2 factors using 9 distinct experimental conditions. Spherical design was used to allow the experiment to be efficient since only 4 new experimental conditions were to be generated in addition to the data collected in the previous steps. Preview length of 65 sec and match score of 80% was determined to be the optimal combination that minimizes browsing time. The estimated browsing time is 9.63325 minutes and the 95% prediction interval is (8.86378,10.40273).

Introduction

Netflix is a subscription-based TV show and movie streaming service which offers up a wide range of content across various genres. With thousands of titles to choose from, it may be difficult for users to make a decision on what to watch. Long browsing time negatively impacts Netflix since it may lead to users losing their interest in Netflix's streaming services. One way browsing time may be reduced is by creating customized suggestions for every Netflix user.

This experiment will be performed in three phases: Factor Screening, Method of Steepest Descent and Response Optimization. Three potentially important factors that may effect browsing time are tile size, match score and preview length. In the Factor Screening Phase, these factors will be analyzed using 2^3 factorial design to determine which ones significantly influence browsing time. Eight different Tile Size/ Match Score/Preview Length combinations will be assigned to 800 Netflix users with 100 users assigned to each combination. Any factors deemed insignificant in the Factor Screening Phase will be ignored in all subsequent phases of experimentation. The metric of interest in this experiment is the average duration of time Netflix users spend browsing a page as opposed to watching Netflix content. Ideally, browsing time should be a small number to ensure that Netflix users don't get overwhelmed by all the options and ultimately lose interest. Main and interaction effect plots will be produced to visualize Factor Screening Phase findings and to assist in formulating conclusions.

The primary goal of the Method of Steepest Descent Phase is to determine the vicinity of the optimum combination of the factors' levels that will optimize the average browsing time. The number of factors analyzed in this section will depend on the number of influential factors identified in the Factor Screening Phase. The Method of Steepest Decent will be applied to move from the initial region of experimentation with a center point (Preview Length, Match Score) = (110,90) towards the vicinity of the optimum in attempts to find an optimal configuration of influential factors that will minimize expected browsing time. A curvature test will be performed in any intermediate steps to determine whether the current experimental region is in the vicinity of the optimum and to reorient towards the optimum if needed. To determine the direction of the path of steepest descent, the first order regression model will be fit. A plot of average browsing time vs. step number will be used to determine which condition minimizes the average browsing time and to understand where the vicinity of optimum is located. Once the vicinity of optimum is reached, this experimental phase will be followed up by a response surface experiment.

The final phase of the experiment is Response Optimization. The objective of this phase is to fit a full second order response surface model to identify the factor levels in the vicinity of the optimum that minimize expected browsing time. Central Composite Design will be used to investigate 2 factors using $2^2 + 2 * 2 + 1 = 9$ distinct experimental conditions. To ensure that the estimate of the response surface at each condition is equally precise, value $a = \sqrt{K}$ will be chosen. This value of a for the Central Composite Design will ensure that the axial conditions are located at an equal distance from the center point as the factorial conditions. The second order linear regression model will be fit to analyze quadratic effects in addition to main effects and second order interactions of influential factors. Using the β estimates of the second order linear regression, 2D contour plots will be produced to visualize the vicinity of the optimum. The elliptical contours will verify that the vicinity of the optimum was indeed identified. Finally, it will help us to determine the factor levels that minimize expected browsing time.

Phase 1: Factor Screening

The objective of this phase is to identify which factors significantly influence Netflix browsing time. An experiment was performed to test how Netflix browsing time is affected by different levels of three potentially important factors: Tile Size, Match Score and Preview Length. The tile size factor corresponds to the ratio of a tile's height to the overall screen height. Smaller tile size values correspond to a larger number of tiles visible on the screen when Netflix is open. Likewise, larger tile size values correspond to fewer visible tiles on the page. Match score factor is a percentage prediction of how likely a specific Netflix user is to enjoy a movie based on their viewing history. The closer match score is to 100%, the more confident Netflix is that a viewer will like the show. The last factor investigated in this experiment is preview length, the duration (in seconds) of a title's preview. To eliminate any potential effect a type of preview may have on Netflix browsing time, it was chosen to keep preview type constant. A title's trailer was shown to Netflix viewers in every experimental condition. The three factors, their levels and regions of operability are summarized in the table below.

Factor	Low Level	High Level	Region of Operability
Tile Size	0.1	0.3	[0.1, 0.5]
Match Score	80	100	[0,100]
Preview Length	100	120	[30,120]

The metric of interest in this experiment is the average duration of time Netflix users spend browsing a page as opposed to watching Netflix content. The corresponding response variable is time one user spends browsing Netflix content before choosing a title to watch. Ideally, browsing time should be a small number to ensure that Netflix users don't get overwhelmed by all the options and ultimately lose interest. Browsing time observations do not include the amount of time users spend watching previews and only count the time spent scrolling and searching. The response variable for every observation is measured in minutes.

A 2^3 factorial experiment was carried out to investigate the three factors of interest and their influence on Netflix browsing time. The $2^3 = 8$ unique combinations of factor levels produced 8 experimental conditions, each of which was assigned $n = 100$ units. Practically speaking, 8 different Tile Size/Match Score/Preview Length combinations were assigned to 800 Netflix users with 100 users assigned to each combination. An option of 2^2 fractional factorial design with only 4 experimental conditions would make the experiment cheaper and faster to run. However, it was abandoned due to a risk to miss important interactions since fractional factorial design obfuscates interactions between factors and does not allow separate analysis of aliased effects.

A linear regression model with the following linear predictor was fit:

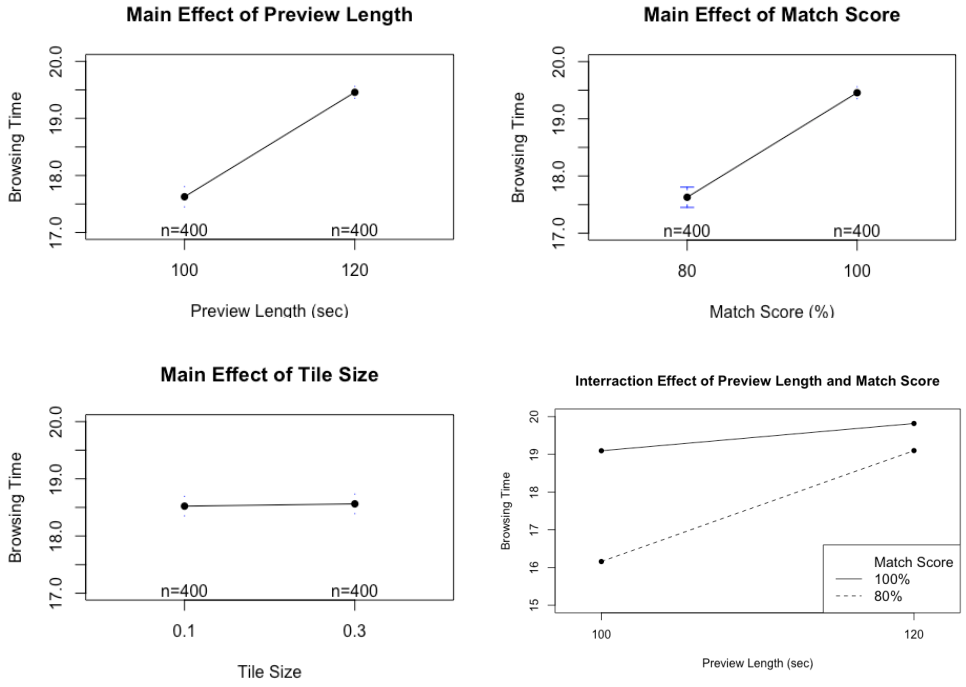
$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3$$

The x_1, x_2, x_3 variables correspond to tile size, match score and preview length respectively. Each of 3 factors is represented by binary variables ± 1, and every experimental condition was identified by a unique combination of ± 1. The significance of main and interaction effects was determined by testing hypotheses that set the relevant $\beta' s$ equal to 0. Since each effect was represented by a single term, the hypotheses of interest involve a single β. Estimation of the $\beta' s$ was carried out by the ordinary least

squares. The following table summarizes β estimates, effects, and p-values of significant main effects and a two-factor interaction. A p-value less than 0.01 was considered statistically significant.

Coefficient	β estimate	Effect	P-value
Preview Length	0.915546	1.831092	$8.329 * 10^{-104}$
Match Score	0.913638	1.827276	$1.742 * 10^{-103}$
Prev. Length:Match Score	-0.554491	-1.108982	$1.599 * 10^{-46}$

The table suggests that preview length, match score, and their interaction are significant. Next, a partial F-test was used to test $H_0 = \beta_1 = \beta_{13} = \beta_{12} = \beta_{123} = 0$. The p-value corresponding to this test was 0.7663 implying that we do not reject H_0. Therefore, the two main effects (preview length and match score) and their two factor interaction are indeed significant. We expect the browsing time to increase by approximately 1.83 minutes when preview length is increased from 100 to 120 seconds or match score increased from 80% to 100%. The main effects of all three factors and interaction effect for preview length and match score are shown on the plots below.



The main effect plots demonstrate that a lower preview length and a lower match score significantly decrease the average browsing time. Furthermore, the interaction plot shows that a lower preview length is especially effective at decreasing browsing time when it's combined with a lower match score. We also observe that the main effect of tile size is very little and can be ignored in all subsequent phases of experimentation. These plots are a great visual proof of the insights we determined using hypothesis testing. In the context of the 8 experimental conditions, the best two average browsing time occurred when both preview length and match score were at their low levels. The average browsing time when tile size has ratio 0.1 and both preview length and match score are at their low levels is 16.13 min and 16.19 min when tile size has ratio 0.3. Changing any of the significant main factors to a high level will increase browsing time.

Phase 2: Method of Steepest Descent

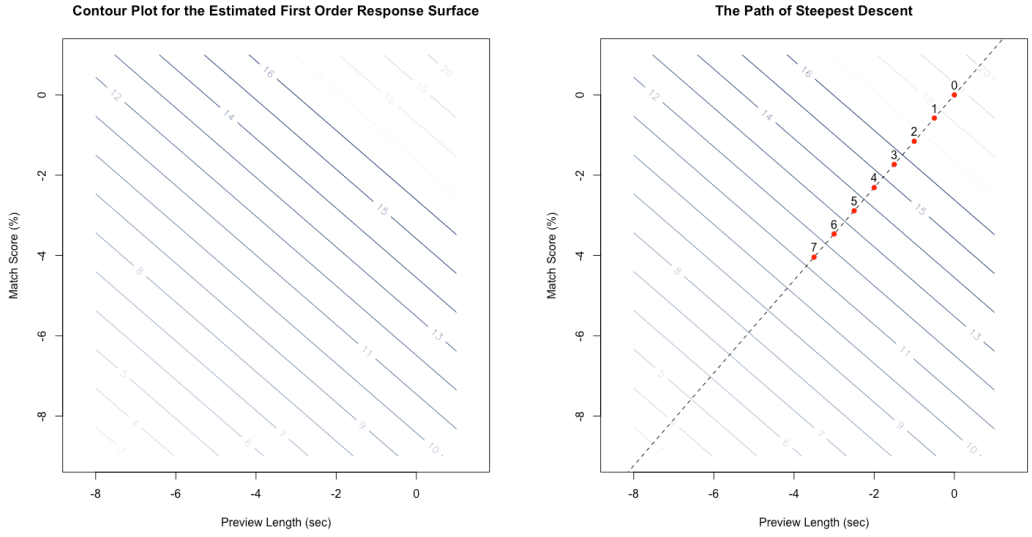
The primary objective of this experiment phase is to determine which combination of the factors' levels will optimize the average browsing time. The method of steepest decent was used to move from the initial region of experimentation toward the vicinity of the optimum and to locate optimal settings of the factors that were identified to be significant in the factor screening phase. We consider 2 design factors which were determined to be influential: preview length and match score. Tile size factor is now kept constant at 0.2 ratio of a tile's height to the overall screen height.

Note that there is no need to perform a curvature test to determine whether the initial experimental region is in the vicinity of the optimum. Therefore, it is necessary to embark down the path of steepest descent. The average browsing times by condition in the 2^2 + Center Point factorial experiment are summarized in a table below.

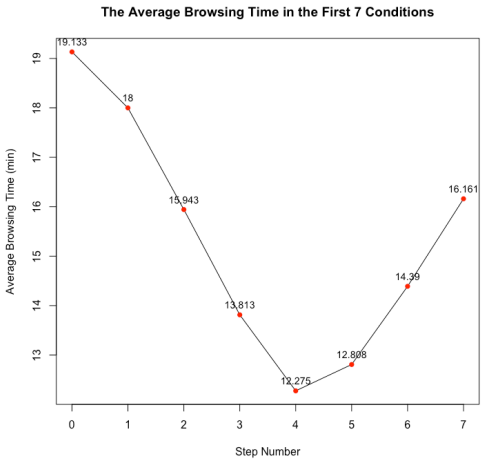
Preview Length	x_1	Match Score	x_2	Average Browsing Time (min)
100	-1	80	-1	16.08378
120	+1	80	-1	19.04088
100	-1	100	+1	19.31789
120	+1	100	+1	19.93944
110	0	90	0	19.13265

The $2^2 + 1 = 5$ unique combinations of factor levels produced 5 experimental conditions, each of which was assigned $n = 100$ units. To determine the direction of the path of steepest descent, the first order regression model with linear predictor $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ was fit using the data described above. The $\widehat{\beta}$ estimates are used to calculate the gradient $g = [\widehat{\beta}_1 \quad \widehat{\beta}_2]^T = [0.895 \ 1.033]^T$

The contour plot for the estimated first order response surface and the path of steepest descent are illustrated on the plots below. The red dots on the dashed line of steepest decent correspond to the experimental conditions conducted along this path, beginning from the center point $(x_1, x_2) = (0,0)$. Step size of $\lambda = \dfrac{0.5}{|0.8946|} = 0.559$ was used, where the value 0.5 was chosen to ensure steps of 5 seconds in preview length.



It is clear from the right plot that moving towards the bottom left corner of the contour plot will cause the average browsing time to decrease. To save on costs and ensure efficiency of experimentation, a total number of 7 steps was taken down the path of steepest descent. This number was chosen to assure that the final step is located approximately 3 coded units away from the center point. This decision allows to explore a big section of the map without exceeding allocated costs for the experiment. The browsing time data for preview length of 110 sec and match score of 90% was taken from a sample generated earlier to save on the experimental costs. The average browsing time and locations in natural units for each of the observed conditions are summarized in a plot and table below.



Step	Preview Length	Match Score	Average Browsing Time
0	110	90	19.133
1	105	84	18
2	100	78	15.943
3	95	73	13.813
4	90	67	12.275
5	85	61	12.808
6	80	55	14.390
7	75	50	16.161

It is clear that Step 4 corresponds to the lowest browsing time. It also turns out that Step 7 is three steps away from the minimum value of average browsing time. Therefore, performing 5 or 6 steps would still help us find the lowest average browsing time but would also increase efficiency of the experiment.

Next, a test of curvature needs to be performed in the region with center point (Preview Length, Match Score) = (90, 67) to determine whether the vicinity of the optimum was reached. The factor levels in coded and natural units for the new experiment are shown in the table below.

Preview Length	x_1	Match Score	x_2	Average Browsing Time (min)
80	-1	57	-1	13.79337
100	+1	57	-1	14.77968
80	-1	77	+1	10.88994
100	+1	77	+1	12.13804
90	0	67	0	15.70727

We perform a 2^2 factorial experiment with a centre point condition to estimate a first-order response surface $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{PQ} x_{PQ}$. The x_1 and x_2 variables correspond to preview length and match score respectively. The x_{PQ} variable is a binary indicator that indicates whether an observation came from one of the factorial conditions or the centre point condition.

The corresponding p-value for $H_0 : \beta_{PQ} = 0$ is $1.546 * 10^{-39}$ indicating that the null hypothesis is rejected, and there is a significant quadratic curvature in this region of the response surface. Therefore, we have arrived in the vicinity of the optimum. This experimental phase should now be followed up by a response surface experiment so that a full second order model can be fit and the optimum identified.

Phase 3: Response Optimization

The objective of this phase is to be able to fit a full second order response surface model to identify the factor levels that minimize expected browsing time. This requires estimating 6 different β coefficients. Central Composite Design will be used to investigate 2 factors using 9 distinct experimental conditions: 4 factorial conditions, 4 axial conditions and 1 center point condition. As estimated by the factor screening, the most influential factors are preview length and match score. According to the Method of Steepest Descent, the optimal preview length is in the vicinity of 90 sec and the optimal match score is in the vicinity of 67%. The low and high factors were kept the same as in the previous section to save on producing new experimental conditions and perform the experiment at higher efficiency.

To ensure that the estimate of the response surface at each condition is equally precise, value $a = \sqrt{2} \approx 1.4$ was chosen. In this design, the axial conditions are set at an equal distance of 1.4 coded units from the center point (Preview Length, Match Score) = (90, 67). This choice of a also allows the experiment to be efficient since only 4 new experimental conditions need to be generated in addition to the data collected in the previous step. The rotatability and the small number of required experiments makes the choice of $a \approx 1.4$ well suited for estimating the coefficients in a second order model. The experimental conditions in both coded and natural units and the average browsing time for each condition are summarized in the table below.

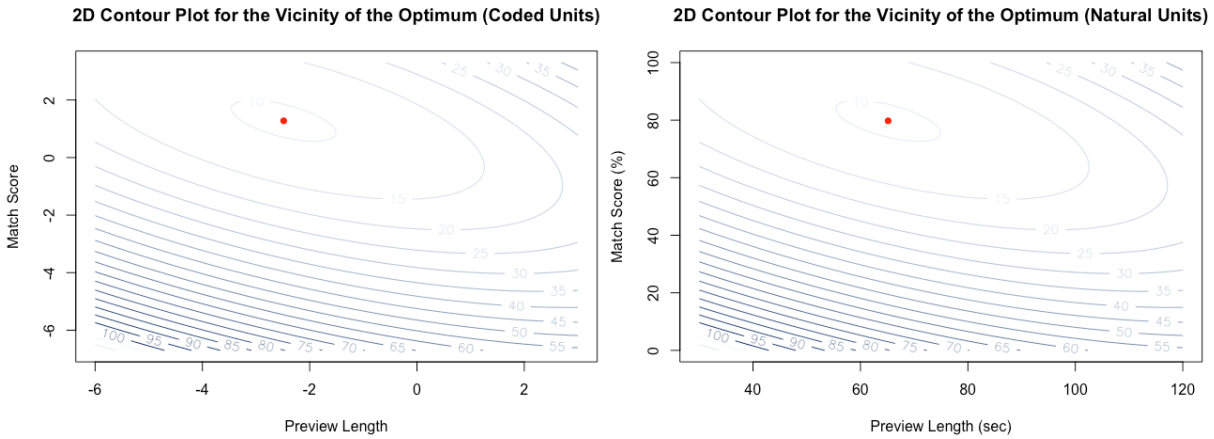
Preview Length	x_1	Match Score	x_2	Average Browsing Time (min)
80	-1	57	-1	13.79337
100	+1	57	-1	14.77968
80	-1	77	+1	10.88994
100	+1	77	+1	12.13804
90	0	67	0	15.70727
104	+1.4	67	0	13.50878
76	-1.4	67	0	13.64516
90	0	81	+1.4	12.53340
90	0	53	-1.4	13.46741

Observe that only the last 4 conditions are different from conditions that were already seen in this experiment. The 9 unique combinations of factor levels produced 9 experimental conditions, each of which was assigned $n = 100$ Netflix users.

The second order linear regression model was fit to estimate β coefficients:

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

The x_1 and x_2 variables correspond to preview length and match score respectively. This allows analysis of quadratic effects in addition to main effects and two factor interaction of preview length and match score. Using this information, 2D contour plots of the fitted response surface were produced. The 2D contour plots in coded and natural units are shown in figures below. The elliptical contours verify that we found the vicinity of the optimum. The stationary point is identified in red on both plots below.



The stationary point for the second order model in coded units is located at $x_1 = -2.484808$ and $x_2 = 1.276511$. In the natural units these values correspond to a preview length of 65.15192 sec and match score of 79.76511%. The expected browsing time at this point is 9.63288 minutes and 95% prediction interval at this optimum is (8.87859,10.38710). For the experimentation to provide a practically feasible optimum, we require preview length to be a multiple of 5 seconds and match score to be an integer. A slightly less optimal combination of the two factors would be a preview length of 65 sec and match score of 80%. In coded units, this combination is located at $x_1 = -2.5$ and $x_2 = 1.3$. The estimated browsing time with 65 sec preview length and 80% match score is 9.63325 minutes. The 95% prediction interval is (8.86378,10.40273).

Note that the estimated browsing time does not significantly increase when we pick a 65 sec preview length and 80% match score combination over the optimum point. Therefore, preview length of 65 sec and match score of 80% defines the optimal combination. In the future experimentation stages if any, Netflix should move forward with this combination into a confirmation phase.