**An experiment and analysis in order to optimize the Netflix homepage**

By Ajjan Inthiran

Optimizing the www.netflix.com homepage by way of minimizing
browsing time - length of time a user spends browsing
Average browsing time

In this project, I conduct a series of experiments to learn about what influences browsing time and how that may be exploited in order to minimize average browsing time. There are infinitely many things that likely influence the amount of time someone spends browsing Netflix, but just four factors will be explored in this project. Each is related to the "Top Picks For. . . " row of the Netflix homepage. This row contains recommendations algorithmically curated for the specific user.

- Tile Size:

The ratio of a tile's height to the overall screen height. Note the tile's aspect ratio is fixed so changing this factor changes the size of the tile, but not its shape. Smaller values correspond to a larger number of tiles visible on the screen, and larger values correspond to fewer visible tiles.

- Match Score:

A prediction of how much you will enjoy watching the show or movie, based on your viewing history. This is recorded as a percentage, with larger values indicating a higher likelihood of enjoyment.

- Preview Length:

The duration (in seconds) of a show or movie's preview.

- Preview Type:

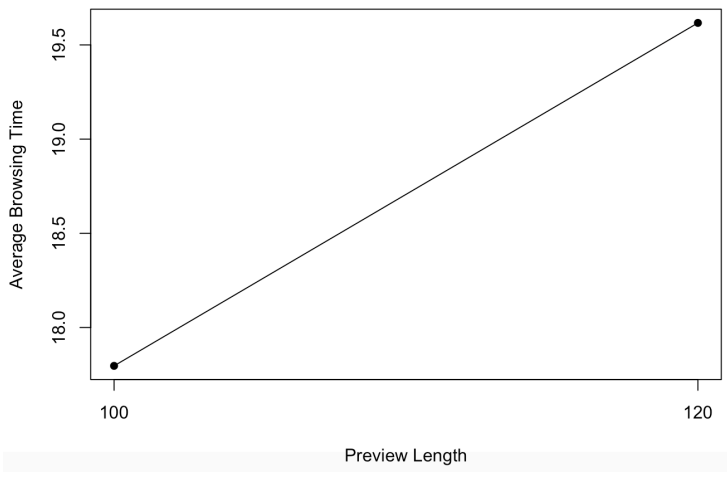
The type of preview that is autoplayed.

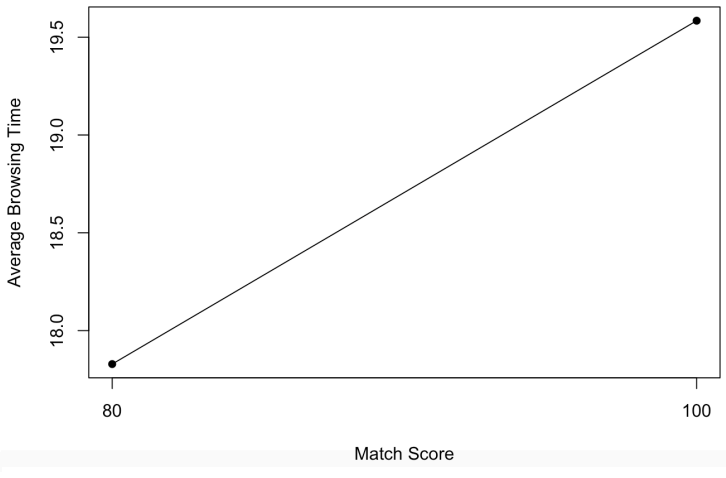
Phase 1
Factor Screening

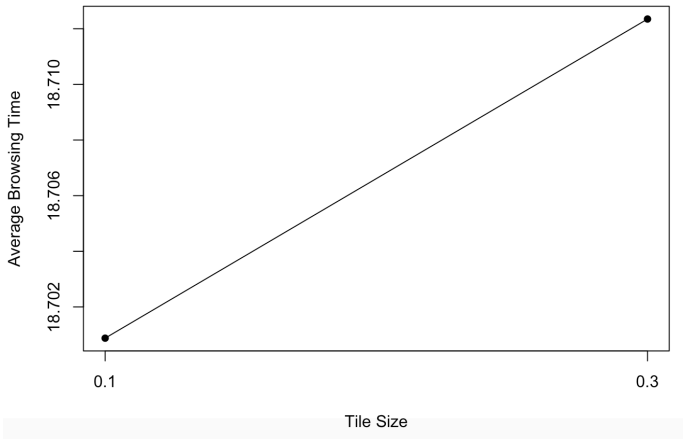
**Question:** Which of the factors, between Tile Size, Match Score and Preview Length significantly influences the response variable, by minimizing the average browsing time of a user on Netflix.
**Plan:** The metric of interest is browsing time. The response variable is average browsing time. The design factors are Tile Size, Match Score and Preview Length. The experimental units are Netflix users. A 2^k factorial design was conducted. The reason for choosing a 2^k experiment over a 2^(k-p) experiment was because there were fewer higher order interaction terms. Thus we would be conducting few experiments on high order interaction terms, so 2^k factorial design would result in less waste, increased accuracy and would be more organized.
**Data:** Particularly a 2^3 design matrix was submitted into a response surface simulator. Upon submission, observations were received with random assignment of n=100 users for each of the experimental conditions. As for our three design factors, Tile Size is the ratio of a tile's height to the overall screen height. Match Score is a prediction of how much you will enjoy watching the show or movie, based on your viewing history. This is recorded as a percentage, with larger values indicating a higher likelihood of enjoyment. Finally, Preview Length is the duration (in seconds) of a show or movie's preview.
**Analysis:** In order to determine which main effects are significant we fit a (linear) regression model with linear predictor given by: B0 + B1x1 + B2x2 + B3x3 , where:
X1 = 1 when the level of the tile size is 0.3, and 0 otherwise.
X2 = 1 when the level of the match score is 100, and 0 otherwise.
X3 = 1 when the level of the previous length is 120, and 0 otherwise.
Because each main effect is represented by just one term, the reduced model in each case arises by setting the appropriate β equal to 0. The reduced models would then each be compared to the full main effects model by way of a partial F-tests with test statistic and p-value calculated as follows, where T ~ F(1, n-p-1)

$$t = \frac{(SSE_{\text{red}} - SSE_{\text{full}})/l}{SSE_{\text{full}}/(n-p-1)}$$

$$p\text{-value} = \mathrm{P}(T \geq t)$$






| Coefficients | Estimate | Standard Error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| Intercept | -6.376e+01 | 7.832e+00 | -8.141 | 1.52e-15 |
| Preview Length | 6.839e-01 | 7.091e-02 | 9.644 | < 2e-16 |
| Match Score | 7.960e-01 | 8.649e-02 | 9.203 | < 2e-16 |
| Tile Size | 1.096e+01 | 3.503e+01 | 0.313 | 0.754 |
| Prev.Length:Match.Score | -6.506e-03 | 7.831e-04 | -8.308 | 4.19e-16 |
| Preview Length:Tile Size | -1.296e-01 | 3.171e-01 | -0.409 | 0.683 |
| Match Score: Tile Size | -7.657e-02 | 3.868e-01 | -0.198 | 0.843 |
| Match Score: Preview Length: Tile Size | 1.034e-03 | 3.502e-03 | 0.295 | 0.768 |

Conclusion:
The output above provides p-values associated with t-tests of the hypothesis
H0:β = 0 vs. HA:β =/= 0 for each of the β's in the model.
As such, to determine whether a given main effect is significant, we need only look at its p-value in the output above. Based on these p-values, only the design factors Prev.Length and Match.Score are statistically significant at the 1% level of significance. Thus Tile.Size, the factor deemed insignificant can be ignored in all subsequent phases. As well, the main effect plots confirmed our results from the summary table. The main effects of Match Score and Preview Length were significant enough to influence average browsing time. The plot of the Tile Size main effect is further evidence of its statistical insignificance.
It is also apparent that the interaction effect of higher Match Scores with shorter Preview Lengths led to lower average browsing times. Consequently, we see that the p-values associated with the Prev.Length:Match.Score interaction are less than 0.01 and that no other two-factor interaction effect has a p-value this small. Thus, Prev.Length:Match.Score is the only two-factor interaction significant at the 1% level.

Introduction

Netflix is one of the fastest growing companies, and many of their such messes can be attributed to their advanced position in managing data and experimental design. Netflix is an online streaming service, with a broad catalogue of movies, original programs, television shows and documentaries. Netflix's commitment to data driven analytics and designed experiments is commendable, as they have continued to refine and develop their service to better suit the needs of their customers.

For some additional insight into the experiment and problem we are trying to solve, Netflix users navigate through a grid system where movies, shows, etc. are displayed in tiles and rows as per a particular category. As a user hovers over a title, this would enlarge the selection and present a quick preview of the program.

The problem we are trying to solve is a hypothetical problem, as we are looking to optimize the www.netflix.com homepage by way of minimizing browsing time - length of time a user spends browsing. Considering the fact that Netflix users are often faced with experience choice-overload and decision paralysis, this hinders the user's experience and it is critical for Netflix to advance towards a solution that improves browsing time across all users.
In this project you will conduct a series of experiments to learn what influences browsing time and how that may be exploited in order to minimize average browsing time. There are infinitely many things that likely influence the amount of time someone spends browsing Netflix, but just four factors will be explored in this project.

Each of the fours factors we will be exploring are related to the "Top Picks For… " row of the Netflix homepage. The first of the factors of interest is Tile Size, the ratio of a tile's height to the overall screen height. Smaller values correspond to a larger number of tiles visible on the screen, and larger values correspond to fewer visible tiles.The next factor would be Match Score, which is a prediction of how much you will enjoy watching the show or movie, based on your viewing history. This is recorded as a percentage, with larger values indicating a higher likelihood of enjoyment. Finally, Preview Length is the duration (in seconds) of a show or movie's preview. This row contains recommendations algorithmically curated for the specific user. We conduct a 2^(3) factorial design since the Prev.Type factor may be ignored for the project.

To provide a deeper understanding of Response Surface Methodology, it explores the relationships between several explanatory variables and one or more response variables. A

critical component of RSM is to use a sequence of designed experiments to obtain an optimal response. Considering this model is only an approximation, this model is easier to estimate and apply throughout the process of designed experimentation.

Considering effective experimentation is sequential, we would like to be well informed as we conduct future experiments. Thus the information attained in one experiment is beneficial in future experimentation, and this is the basis of response surface methodology (RSM). At one point, we only used screening experiments to identify which among several factors significantly influence the response variable. We follow up on the ideas of screening experiments by now conducting further screening experiments where the primary goal is response optimization.We use the method of steepest ascent/descent and response surface designs to locate optimal settings of the factors that were identified as significant in the screening phase.

As it pertains to the goals of Response Surface Methodology, this application has a variety of purposes and objectives. Of these objectives, this includes generating knowledge in the experimental domain of interest, reliably estimating the experimental variability or the pure error, guaranteeing the adequacy between the proposed model and the experimental data. As well, the RSM can be used when detecting lack of fit, predicting the observed response, as exactly and precisely as possible, in points within the experimental domain where no experiments were done and when trying to identify outlier data more easily. The final set of purposes of the Response Surface Methodology that I would like to highlight include proposing sequential strategies to carry out the experimentation with potential alternatives as per the results obtained and finally, maintaining a high efficiency with respect to use of resources. This accounts for time, economical costs, and other possible limitations.

RSM can be tremendously applicable useful in a variety of real world applications revolving around designed experimentations and data driven analytics. To make the decision making possible under uncertainty conditions, RSM reduces the ambiguity. As we develop the response surface of $y$, we inch our processes closer towards the optimum, considering any constraints. This opens up possibilities for better designs as we move towards the optimum, that more closely approximate conditions at the optimum. Hence, its sequential nature.

Phase 2 - Method of Steepest Descent

**Question:** Perform a method of steepest descent analysis to move from the initial region of experimentation toward the vicinity of the optimal average browsing time.

**Plan:** We focus on the Preview Length factor and a Match Score factor due to their significance from the factor screening component. The initial region of experimentation is not in the vicinity of the optimum, and embarking down the path of steepest descent is necessary. You may use this fact without justification. We begin with a 2^2 factorial experiment with a centre point condition. Upon determining the initial gradient, a traversal down the gradient at step size, 5 seconds, where each centre point on the line was inputted into the response surface simulator.

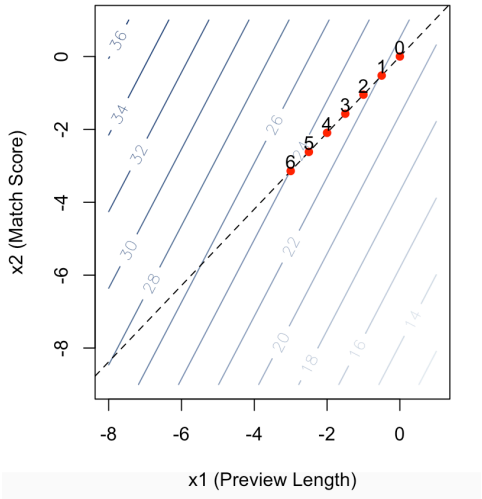
**Design:** To begin the method of steepest descent procedure we use the aforementioned data to fit the first order regression model with linear predictor: n = B0 + B1x1 + B2x2. This was used to determine the direction of the path of steepest descent. The beta values calculated are shown in the model summary and were applied to generate the gradient.

Model Summary is shown:

| Coefficients | Estimate | Standard Error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| Intercept | 18.79085 | 0.03466 | 542.21 | < 2e-16 |
| Preview Length | 0.90836 | 0.04244 | 21.40 | < 2e-16 |
| Match Score | 0.95133 | 0.04244 | 22.41 | < 2e-16 |

The plot below depicts the contours of the estimated first order response surface:

$$\mathbf{g} = \begin{bmatrix} \hat{\beta}_1 & \hat{\beta}_2 \end{bmatrix}^T$$



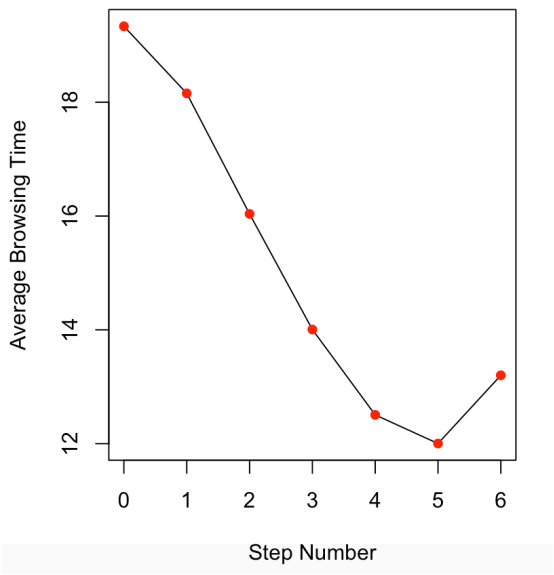

Gradient Calculated: [0.9083587, 0.9513292]

**Analysis:** This path of steepest descent is depicted by the dashed black line in the plot. The red dots signify the experimental conditions conducted along this path, beginning from the centre point (x1, x2) = (0, 0).

The step 0 point which we start with, and as you may see the direction in which the points take follows the gradient. We fit the first order model to determine the direction of the path of steepest descent. The dashed black line represents the gradient. Each of the points represent each of the steps we take along the path.

Step Size: Lambda = (Change in x1)/|B1| = 0.5/|0.9083587|

Numerator value was chosen to ensure steps of 5 seconds in Preview Lengths. We were able to generate a numerator that is applicable in this instance by applying R code that ensures steps of 5 seconds for Preview Lengths. We chose a step size of 5 seconds, firstly because a larger step may lead to inaccuracies when reading the path. Since




Prev.Length had to be a factor of 5, steps of 5 seconds would ensure this requirement, as we conduct a full exploration of the path.

By what we have plotted, Average Browsing Time proceeds to increase at step 6 after having decreased from steps 0 to 5. At step 6 we stopped seeing incremental improvements in our MOI. Thus, step 5 is the condition that minimizes Average Browsing Time and wherever we are at step 5 is probably the vicinity of the optimum. We can draw the conclusion that proceeding further with the gradient would stray away from the optimum.

We follow this with a test of curvature around this area to determine whether we have reached the optimum. We must re-code what we consider low vs. high. We should follow this up with 2^2 factorial conditions to ensure we're close to optimum. We will re-center our coded scale in this new region as follows: Preview Length: 70 vs 100, Match Score: 50 vs 80.

Fit a linear regression model, linear predictor: n = B0 + B1X1 + B2X2 + B12X1X2 + Bpq Xpq

The result of interest are presented in the summary chart below.

| Coefficients | Estimate | Standard Error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| Intercept | 11.97404 | 0.10223 | 117.13 | <2e-16 |
| Preview Length | 1.03829 | 0.05111 | 20.31 | < 2e-16 |
| Match Score | -1.35818 | 0.05111 | -26.57 | < 2e-16 |
| xPQ | 2.73039 | 0.11429 | 23.89 | < 2e-16 |
| Prev.Length:Match.Score | 1.73132 | 0.05111 | 33.87 | < 2e-16 |

Conclusion: As per the final linear model constructed, we direct our attention to the row of the pure quadratic. It is statistically significant, thus leading to the belief that this area is in the vicinity of the optimum. We may reject the null hypothesis that the beta of the pure quadratic is equivalent to 0. There is significant quadratic curvature in this region of the response surface. Now we should commence phase 3 and perform a response surface design and fit a full second order model. Based on the final linear model constructed since all coefficients were significant, I knew to stop.

Phase 3

Response Optimization

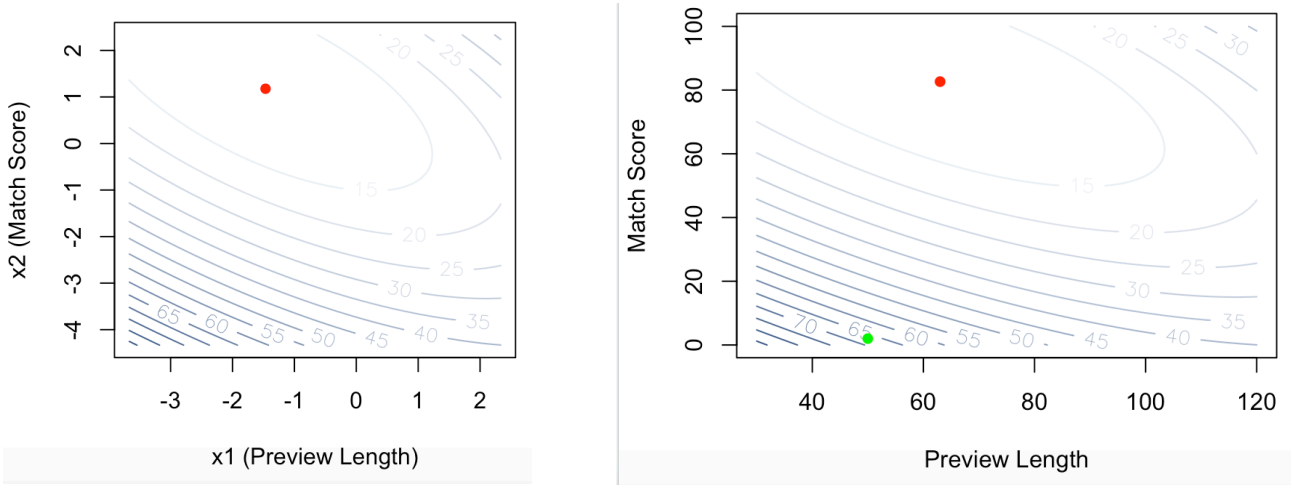
**Question:** The objective of this experiment is to fit a second order model with the data. Consequently, this would dictate the appropriate values of Match Score and Preview Length, thus optimizing the average browsing time amongst Netflix users. As for this sequence of the experiment, we maintain consistency by applying the central composite design on the 2^2 factorial design from earlier portions.

**Plan:** The design used was a central composite design on the full $2^2$ factorial design that was explored earlier. The process of the response optimization is a logistical regression process, as schwa make certain considerations. The factorial conditions of Preview Length was tested at high level of 100 and low level of 70 seconds, and Match Score was tested at high level of 80% and low level of 50%. These “high” and “low” values were chosen based on what was seen in the path of steepest descent. Ideally, we would like to choose low and high values wide enough to try to include that quadratic curvature. That way the test for curvature will be well-informed and the follow-up CCD will be well-positioned.The high and low values were chosen to maintain a level of consistency among optimal regions, the initial point in the steepest descent process had bounds of [100,120] and [80,100] for Preview Length and Match Score.To find optimal values of these factors a follow-up two-factor central composite design was run in order to fit a second-order response surface model.

**Design:** Given that the current central point was not on the bounds of the high and low values, we must consider axial conditions. The axial conditions be with a = root(2). To follow, n=100 users were then randomized into m=9 conditions, so that the estimate for the response surface at each condition is consistent in terms of accuracy. A table of the experimental conditions submitted to the simulator is shown below. We bind the output with the final dataset we had from the method of steepest descent, to conduct the he full 9-condition CCD that is required in Phase III.

| Prev.Length | Match.Score |
|---|---|
| 65 | 65 |
| 105 | 65 |
| 85 | 45 |
| 85 | 85 |

**Analysis:** Contour Plots of fitted response surface are shown below in Coded and Natural Units:




The output from the fitted second order logistic regression model is present below:

| Coefficients | Estimate | Standard Error | t-value | Pr(>\|t\|) |
|---|---|---|---|---|
| Intercept | 11.99083 | 0.10033 | 119.52 | <2e-16 |
| X1 | 1.17838 | 0.03700 | 31.85 | < 2e-16 |
| X2 | -1.34590 | 0.03700 | -36.38 | < 2e-16 |
| I(x1^2) | 1.09652 | 0.06310 | 17.38 | < 2e-16 |
| I(x2^2) | 1.65066 | 0.06310 | 26.16 | < 2e-16 |
| X1:x2 | 1.73132 | 0.05085 | 34.05 | < 2e-16 |

**Conclusion:**

In Natural Units, this optimal is located at a Preview Length of 63.00658 Seconds and a Match Score of 82.64929%.

95% prediction interval at this optimum: (10.0249325120862,10.6453392444549)

95% prediction interval at convenient near-optimum: (14.8516780199766,15.1231073726637)

The 95% Confidence interval we determine is (14.8516780199766,15.1231073726637)

## Executive Summary

The problem we are trying to solve is a hypothetical problem, as we are looking to optimize the www.netflix.com homepage by way of minimizing browsing time - length of time a user spends browsing. Considering the fact that Netflix users are often faced with experience choice-overload and decision paralysis, this hinders the user's experience and it is critical for Netflix to advance towards a solution that improves browsing time across all users.

In this project we complete a series of experiments to learn what influences browsing time and how that may be exploited in order to minimize average browsing time. There are infinitely many things that likely influence the amount of time someone spends browsing Netflix, but just four factors will be explored in this project. Over the course of this design experimentation, we conducted a factor screening experiment, method of steepest descent analysis, and a central composite design, response optimization. These types of experimentation was undertaken to realize which of the design factors would be most significant, as well as, where the vicinity of the optimum factor levels lie. Over the course of the experiment, we would submit certain design matrices into a response surface simulator that further generated n=100 random users for each of the conditions, but also providing average browsing time for each condition.

From the factor screening experiment, we conclude that the main effects of Match Score and Preview Length were significant enough to influence average browsing time. Tile Size was statistically insignificant.

In the method of steepest descent, we re-center our coded scale in this new region to ensure that we are close to the optimum with Preview Length: 70 to 100 and Match Score: 50 to 80. As per the final linear model constructed, we conclude the row of the pure quadratic is statistically significant. Thus leading to the belief that this area is in the vicinity of the optimum. As for our final design experimentation, we proceed with the response optimization. After our analysis, in Natural Units, the optimal is located at a Preview Length of 63.00658 Seconds and a Match Score of 82.64929%.

The objective of conducting this factorial experiment as well as a response surface methodology experiment was to develop a deeper understanding of design experimentation, the sequential nature of experiments and the necessary presentation elements of experimental design.