Executive Summary

To help Netflix users spend less time browsing and more time watching, I conducted a series of experiments manipulating tile size, preview length, preview type, and match score as factors to find the optimal combination that would minimize average browsing time. Starting a series of experiments by looking at the extremes enabled eventual convergence toward an optimum. The optimum location was at a preview length of 70 seconds, a match score of 71%, a tile size of 0.2 (default value), and a teaser/trailer preview type. The average browsing time at this location was estimated at 10.47 minutes, and I can report with 95% confidence that the expected browsing time at this location is between 10.26 and 10.68 minutes.

Introduction

Have you ever found yourself scrolling through Netflix to find a movie or TV show to watch, but you just can't seem to find what you are looking for? This is not an uncommon phenomenon, and it is often referred to as decision paralysis. In efforts to minimize such decision paralysis, experiments were conducted for this report to find the configuration of Netflix's 'top picks for you' row that would minimize the user's average browsing time. Many factors may influence this metric of interest, but for this report, I am only going to focus on four: Tile Size (the ratio of a tile's height to the overall screen height; default value: 0.2), Preview Length (the duration of a show or movie's preview; default value: 75 seconds), Preview Type (either a teaser/trailer or actual content; default: teaser/trailer), and Match Score (a prediction of how much you will enjoy watching the show or movie based on previous viewing history; default: 95%).

After collecting enough evidence to determine that Tile Size was not a significant factor, I started with extreme values for each of the other factors and iteratively narrowed down the search region. In the remainder of this report, I will walk through the different experiments that I decided to run, explaining why I chose to run and how I executed each of them.

Experiments and Findings

Experiment 1: Is tile size a significant factor?

To limit the search space, I first decided to determine whether Tile Size (TS) is a significant factor, as the significance of this feature is questionable. I ran a factorial experiment on collected data using the default level for Preview Length (PL), Match Score (MS), and Preview Type (PT), and used two most extreme levels for Tile Size (TS) {0.1, 0.5}. This yielded two different experimental conditions (n = 100 experimental units in each condition). After checking that the data in each experimental condition were normally distributed using a Kolmogorov-Smirnov test and that both conditions had the same variance using an F-test, I employed a Student's t-test to compare their average browsing times (ABTs). Results of this test (t = 0.293; p-value = 0.770) suggested that there is no significant difference in ABTs of users assigned to TS 0.1 and 0.5. Therefore, I concluded that this factor is not a significant one — permitting use of its default value (0.2) for subsequent experiments.

Experiment 2: Factor screening through a 2^3 factorial experiment

After determining the lack of significance of questionable features, I wanted to learn more about the potential interaction effects between the factors as well as their main effects.

In order to achieve this, I used a 2^3 factorial experiment: a particular, minimal, factorial experiment in which I only choose two levels per factor. The big question here was how to choose those levels. To define a search space I was confident contained the optimum, I decided on a much safer approach, by choosing the two extreme values for each factor: {30,120} for PL, {0,100} for MS, and {TT, AC} for PT (no other choices for this last factor). After this, I collected n=100 experimental units for each of the eight (2^3) resulting experimental conditions, and ran a linear regression on the data, using all factors and their interactions.

To analyze the results, I first looked at the summary table of the previously fitted model, and noticed that the two-factor PL-to-MS interaction was the only significant interaction effect. To confirm this, I compared the full model to the reduced model containing all main effects and only the two-factor PL-to-MS interaction effect using an ANOVA F-test, and failed to reject H0 (reduced model and full model perform equally well), thus confirming that the other interaction effects were not significant.

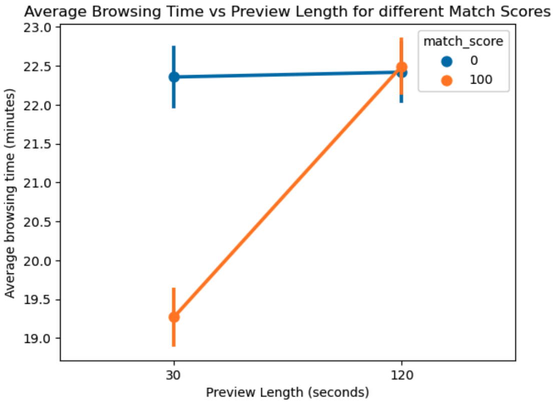


Figure 1

Through analysis of the PL-to-MS interaction plot I observed that, when the MS is 0, the ABT doesn't seem to be very different for PL=30 and PL=120 (see Figure 1). Using this information, I hypothesized that, while the regression analysis suggested that the main effect of preview length was not significant when the levels of match score were {0, 100}, a true significant main effect may exist when data is collected using higher match scores. For this reason, I decided to keep exploring the effects of PL in succeeding experiments.

Because an interaction effect appeared to exist between preview length and match score, I was unable to make inferences about the main effect of those two factors. However, since I concluded that PT had no significant interaction effect with any of the other factors, I was able to study its main effect.

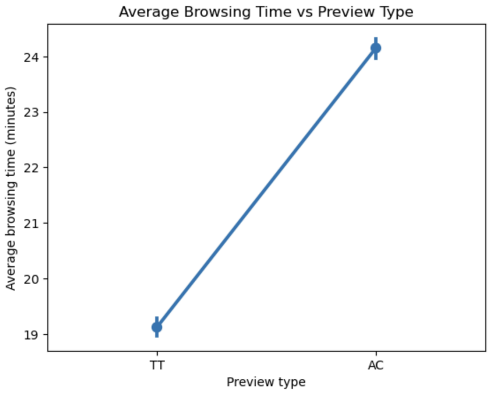
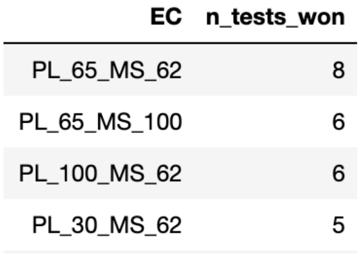


Figure 2

Figure 2 demonstrates that the level of PT associated with the lowest ABT is 'teaser-trailer' (TT). And since, from the summary table of the linear regression model fitted earlier, I concluded that this factor has a significant main effect, we can conclude that the difference observed is significant. Therefore, for all subsequent experiments, I decided to hold the preview type constant as TT. This allowed me to reduce by two the number of experimental conditions needed in all subsequent experiments.

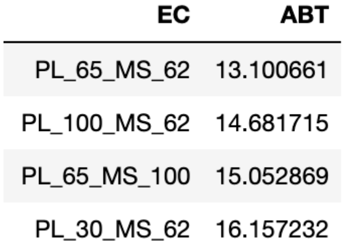
Experiment 3: Second factorial experiment

Given the results of Experiment 2, I decided to add intermediate points for each of the significant continuous factors. I held constant trailer/teaser for preview type and a tile size of 0.2, but experimented with three levels for preview length (30, 65, and 100) and for match score (25, 62, and 100). This yielded nine experimental conditions. I decided to automate and perform pairwise t-tests comparing the ABT of all permutations of the conditions to each other. I then ranked the conditions by the number of tests they "won" (i.e., ABT significantly lower than in the other condition in each pairwise test). This method allowed me to determine the four best experimental conditions:



EC	n_tests_won
PL_65_MS_62	8
PL_65_MS_100	6
PL_100_MS_62	6
PL_30_MS_62	5

Figure 3



EC	ABT
PL_65_MS_62	13.100661
PL_100_MS_62	14.681715
PL_65_MS_100	15.052869
PL_30_MS_62	16.157232

Figure 4

From here, I concluded that the parameters for the optimum condition seem to be a PL between 30 and 100 with more weight to the 65 second condition and a MS between 62 and 100.

Experiment 4: Triangle-shaped experiment

The results of the last experiment allowed me to narrow the search space even further for a fourth factorial experiment. The four best performing points from Experiment 3 created a triangle shape. Using this shape, I calculated the triangle's centroid along with three points equidistant from the centroid and each corner. This left me with four new configurations of PL and MS: PL65MS88, PL65MS75, PL50MS68, PL85MS68. I then collected n=100 experimental units for each of these four new ECs, and computed their ABTs:

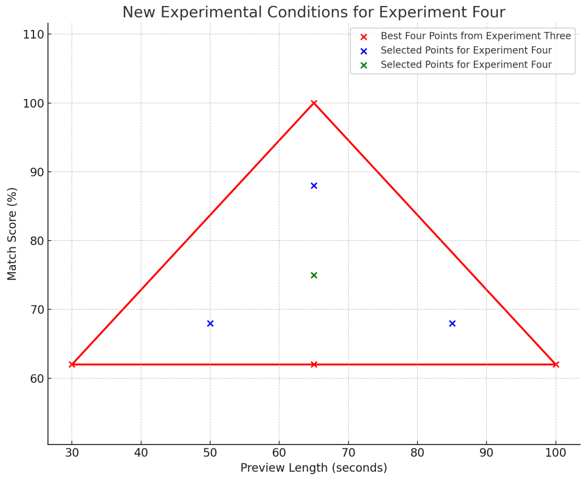
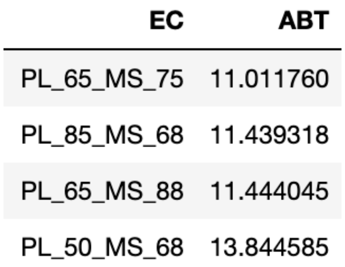


Figure 5

Considering all data collected through Experiments 1 - 4, I observed that PL = 65, MS = 75, PT = TT, and TS = 0.2 produced the lowest observed ABT among all point compared prior to Experiment 4.



EC	ABT
PL_65_MS_75	11.011760
PL_85_MS_68	11.439318
PL_65_MS_88	11.444045
PL_50_MS_68	13.844585

Figure 4

Experiment 5: Central composite design

In my final experiment, I decided to conduct a more thorough exploration of the space immediately surrounding the best point identified in Experiment 4 (PL_65_MS_75). To do so, I selected a grid of the eight points immediately surrounding the hypothesized optimum.

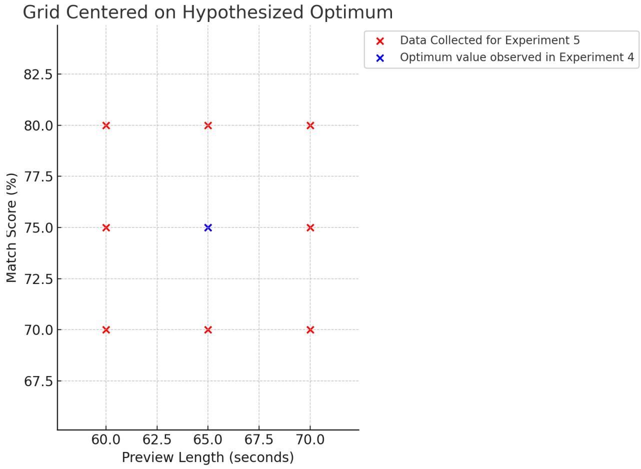


Figure 7

I collected data using these eight new experimental conditions and employed a central composite design (CCD) to generate a second-order model to approximate the true but unknown response surface.

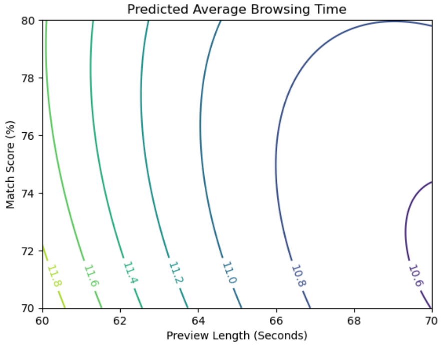


Figure 8

Using this model, I identified a stationary point at PL_70_MS_71 with a corresponding estimated ABT of 10.59 minutes. I then proceeded to collect n = 100 observations at this exact experimental condition. Using this configuration, I observed an ABT of 10.47 minutes – confirming that the estimation of the ABT at this point was close to its true value. Using data collected in Experiment 5, I determined that the point located utilizing CCD is better than the hypothesized optimum from Experiment 4.

Conclusion

Analysis utilizing a series of targeted t-tests, ABTs in each condition, and second-order model approximation allowed me to estimate that a preview-length of 70 seconds and a match score of 71% yielded the shortest ABT. Data collected at the estimated optimum (PL of 70 seconds, a MS of 71%, a TS of 0.2 (default value), and a teaser/trailer preview type) returned the lowest ABT of collected data.

It is, however, important to note a major limitation in the provided analysis: while the estimated optimum returned the smallest ABT *of collected data*, I am not able to confirm that a condition returning a slightly lower ABT does not exist. Indeed, when estimating the response surface in Experiment 5, the estimated optimum fell outside of the search space's window (see Figure 8). While, ideally, I would have liked to approximate a search space containing a true minimum, I needed to estimate the stationary point as being outside of the available window. I ultimately decided to select this configuration to optimize the experiment's efficiency. However, in future experiments with more resources I would move this search space in the direction of the steepest descent and again utilize a central composite design.

In conclusion, this analysis determined key factors influencing reduced ABT on the Netflix homepage: Preview Type, Preview Length, and Match Score. Notably, analysis demonstrates that Tile Size does not exert a significant impact on ABT, but an optimal configuration of other factors' levels do.