Executive Summary

To help Netflix users spend less time browsing and more time watching, we 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. We started our series of experiments by looking at the extremes, and eventually, we converged towards 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 we are 95% confident 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? Then finally you make a decision not to watch anything? This is not an uncommon phenomenon, and it is referred to as decision paralysis. Netflix would rather have you actually watch something than not watch anything at all, which is why they have a row specifically curated for the individual user called "Top Picks for…". Our goal is to find the configuration of this row that would minimize the average browsing time (in minutes). Many factors influence this metric of interest, but for this project, we're only going to focus on four of them: 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%).

We have chosen to employ multiple factorial experiments to converge towards the optimum. After having shown that Tile Size wasn't a significant factor, we started off with extreme values for each of the remaining factors and iteratively narrowed down the search region. In the remainder of this report, we will walk through the different experiments that we have decided to run, explaining why we have chosen to run and how we executed each of them.

The Experiments

Experiment 1: Is Tile Size a significant factor?

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

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

The second step in our experimental journey was to perform a factor screening to learn more about the potential interaction effects between the factors as well as their main effects.

In order to achieve this, we used a 2^3 factorial experiment: a particular, minimal, factorial experiment in which we only choose two levels per factor. The big question here was how to choose those levels. Initially, we wanted to decide based on intuition, for instance ruling out values of MS that were below a certain threshold (around 20%). But we quickly realized that this approach was not right. Indeed, prior to experimenting, we don't know how Netflix users behave, and so we can't assume anything about it. Plus, it was very risky because had our intuition turned out to be incorrect, we could have missed the optimum from the beginning, thus making the rest of the project obsolete. Therefore, we 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, we collected n=100 experimental units for each of the eight (2^3) resulting ECs, and ran a linear regression on this data, using all factors and their interactions.

To analyze the results, we 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, we 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.

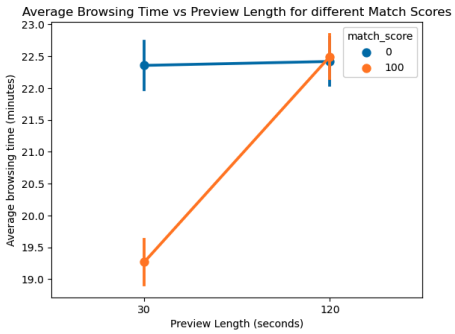

Figure 1

Through analysis of the PL-to-MS interaction plot we observe 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, we hypothesized that, while our 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, we decided to further explore the effect of PL in succeeding experiments. The above plot also confirmed visually the fact that there was a significant PL-to-MS interaction effect. Indeed, the effect of PL on ABT depends on the value of MS.

For this reason, it doesn't make sense to make inferences about the main effect of those two factors. However, since we concluded that PT had no significant interaction effect with any of the other factors, we were able to study its main effect.

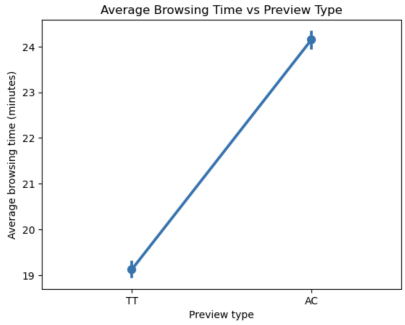

Figure 2

On this plot, we observe that the level of PT associated with the lowest ABT is TT. And since, from the summary table of the linear regression model fitted earlier, we concluded that this factor has a significant main effect, we know that the difference observed here is significant. Therefore, for all subsequent experiments, we have decided to keep PT to TT, which happens to be its default value. This allowed us to reduce by two the number of ECs needed in all subsequent experiments.

Experiment 3: Third factorial experiment

Given the results of our last experiment, we decided to add intermediate points for each of our significant continuous factors. We held constant trailer/teaser for PT and a TS of 0.2, but

experimented with three levels for PL (30, 65, and 100) and for MS (25, 62, and 100). This yielded nine ECs. We decided to automate and perform pairwise t-tests comparing the ABT of all permutations of the conditions to each other. We 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). We found that this ranking method gave similar results as the more "natural" one that consists of ranking the points using their ABT(the lower the better). Here are the four best ECs:

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

Figure 5

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 6

Considering all data collected through Experiments 1 - 4, we observed that PL = 65, MS = 75, PT = TT, and TS = 0.2 produced the lowest observed ABT among all point compared so far

Experiment 5: central composite design

In our final experiment, we 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, we selected a grid of the eight points immediately surrounding the hypothesized optimum.

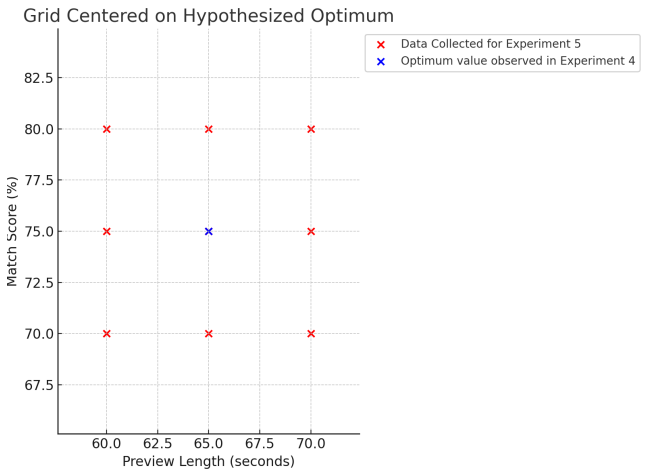

Figure 7

We 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.

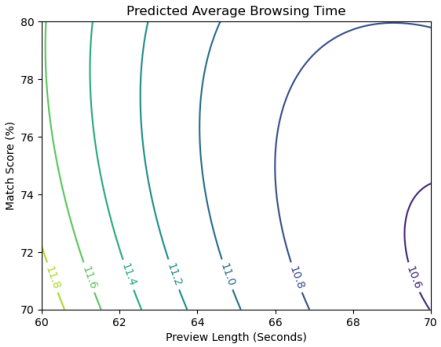

Figure 8

Using this model, we identified a stationary point at PL_70_MS_71 with a corresponding estimated ABT of 10.59 minutes. We then proceeded to collect n=100 observations at this exact EC. Using this configuration, we observed an ABT of 10.47 minutes–thus confirming that our estimation of the ABT at this point was close to its true value. Through Experiment 5, we determine 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 the team to estimate that a PL of 70 seconds and a MS 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 our analysis: while our estimated optimum returned the smallest ABT *of collected data*, we are 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, we would have liked to approximate a search space containing a true minimum, we needed to estimate the stationary point as being outside of our window. We ultimately decided to select this configuration to optimize the experiment's efficiency. However, in future experiments with more resources we would move this search space in the direction of the steepest descent and again utilize a central composite design

In conclusion, our in-depth analysis determined key factors influencing reduced ABT on the Netflix homepage: Preview Type, Preview Length, and Match Score. Notably, we showed that Tile Size did not exert a significant impact on ABT, and found an estimation of the configuration of those factors' levels that yielded the lowest ABT.