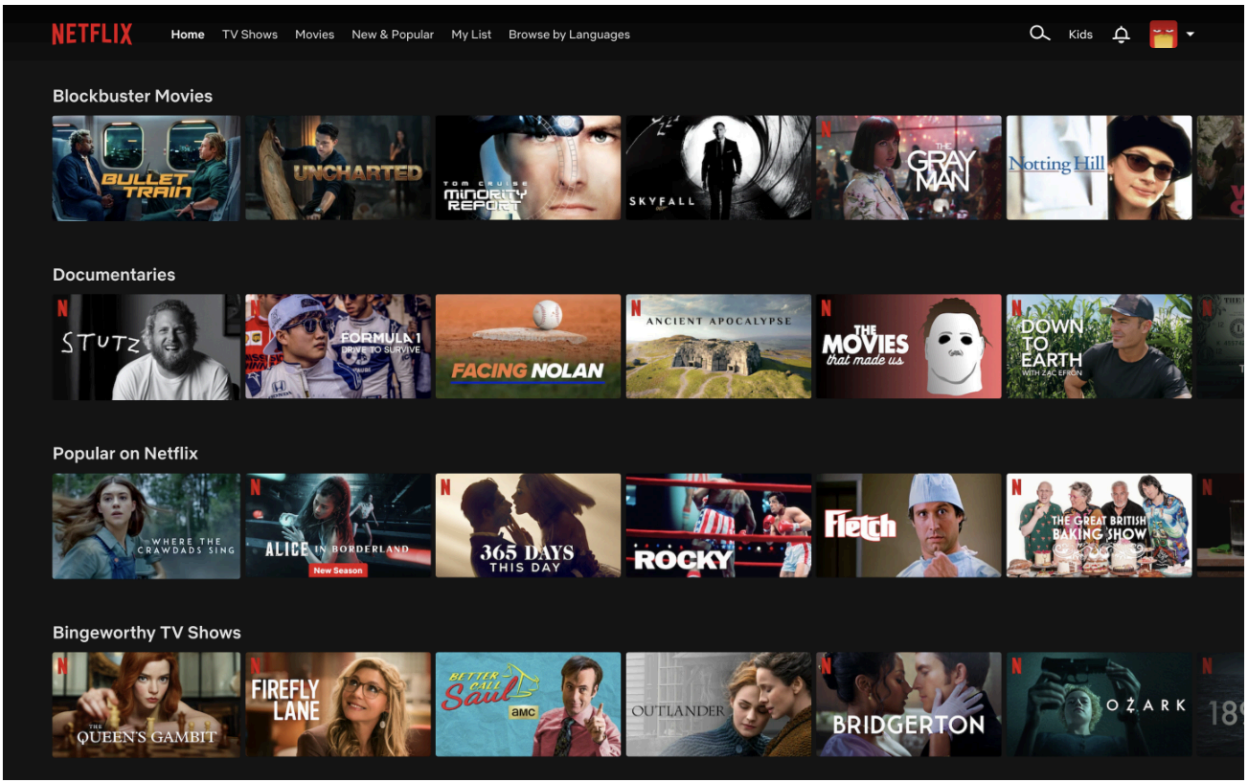
MSDS 629: Final Project

Due: Friday January 19, 2024 by 11:59pm PST



An Experimental Journey:
Netflix-Inspired Simulations for Browsing Efficiency

Group Members:

Brandon Hom

Eren Bardak

Laila Zaidi

Zoe Le

Table of Contents

1. Executive Summary
2. Introduction
3. The Experiments
4. Conclusion

Executive Summary

The goal of this project is to find an optimal set of conditions to minimize a Netflix user's browsing time. We used fundamentals of A/B testing such as factorial design, F-tests, randomized searching, and simulation to gather and analyze browsing time data. We found that the optimal conditions were: Tile Size = 0.2 (default), Match Score = 77%, Preview Length = 75 seconds, and Preview Type = Teaser/Trailer (TT) with a 95% confidence interval of browsing time being about 9.78 to 10.14 minutes.

Introduction

Our goal is to minimize a user's browsing time on Netflix's homepage. A common issue faced by these users is decision paralysis, a state of overchoice leading to longer decision times that potentially reduces content consumption. By optimizing the four conditions Match Score, Tile Size, Preview Length and Preview Type, we are aiming to make the process of choosing content more efficient and less overwhelming for a user.

To effectively address the goal of this experiment, our approach involved gatekeeper tests and randomized searching to gain initial understanding of the data. To ascertain the statistical significance of the various design factors, Analysis of Variance with linear Regression, 2^k tests, and the Partial F-test were utilized. Once significant factors were established, we used the metric of interest (average browsing time) to discern the most optimal set of factors and levels that minimize browsing time.

Here is an outline of the remainder of this report:

- **The Experiments:** Walkthrough of our experimental design, execution, and analysis. This section details how each experiment was set up, the rationale behind our choices, and the insights gained from these simulations.
- **Conclusion:** We summarize the findings of our experiments and simulations. We also reflect on the limitations of our study.

The Experiments

Our experimental journey took place over five steps.

Step One: Determining Significant Factors

To gain insights into the factors influencing the response, we performed a 2^4 factorial experiment. The levels we chose to experiment with were Tile_Size(0.1, 0.5), Prev_Length(30, 120), Match_Score(10,100) and Prev_Type(TT, AC). Then using the simulator we created 100 samples for each condition. We applied a regression to the data, focusing on the significance of main effects and two-way interaction effects, which are summarized in the table below.

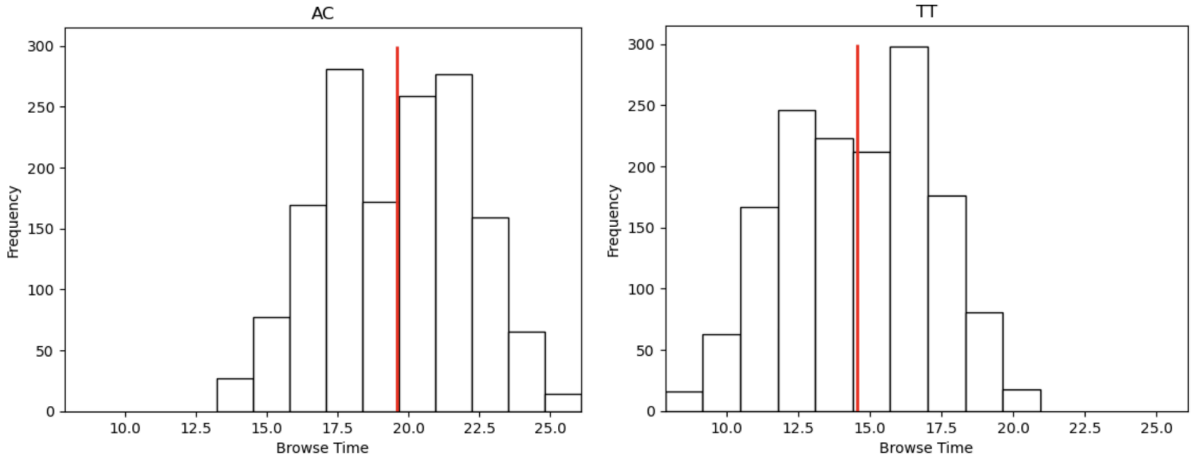
At a 5% significance level, Match Score and Preview Type exhibit a significant main effect with p-values being very close to zero. Although Preview Length does not have a significant main effect, it does have a significant interaction effect with Match Score. Tile Size did not have any significant main effects or interaction effects. Given these results of the 2^4 factorial experiment, we decided to ignore Tile Size and further explore the other three factors—Preview Length, Match Score, and Preview Type—in the next set of experiments.

Factors	t	p-value
Tile_Size	-1.238	0.216
Prev_Length	-0.915	0.360
Match_Score	-22.225	0.000
Prev_Type	-36.568	0.000
Tile_Size: Prev_Length	0.887	0.375
Tile_Size: Match_Score	0.851	0.395
Prev_Length: Match_Score	15.350	0.000
Tile_Size: Prev_Type	0.649	0.517
Prev_Length: Prev_Type	0.683	0.495
Match_Score: Prev_Type	0.209	0.835

Step Two: Determining Optimal Preview Type

We wanted to understand how varying Preview Type would affect browsing time. We created 1500 samples for each preview type (AC and TT) with a set of 30 experiments: Preview Length increased by 15 units with each change in the set of six rows, starting at 45 and reaching up to 105. Match Score was adjusted every 10 rows, starting at 50 and increasing by 20, while Tile Size remained at 0.2 (default level) due to its insignificance across all trials. Each Preview Type, AC and TT, alternated every row, showing a consistent trend where TT resulted in shorter browsing times than AC. Then, we conducted a T-test to test the null hypothesis that

the mean browsing time under the AC condition is greater than or equal to the TT condition with the observed test statistic value “53.8816” and p-value=0m we reject the null hypothesis. Hence, experiments with Preview Type = TT significantly reduce the average browse time over Preview Type=AC. Below is a histogram comparing the browsing times of each Preview Type:



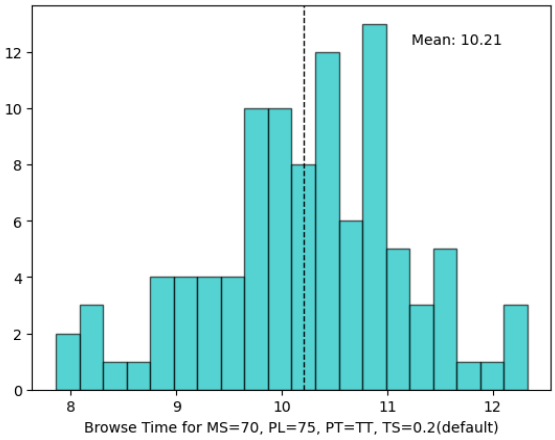
“Mean browsing time with preview type AC is 14.54, and with preview type TT, it is 19.57.”

Step Three: Investigating Optimal Ranges for Preview Length and Match Score

Our next objective was to understand how varying Preview Length and Match Score would affect browsing time. At this point, we decided we want to focus more on accuracy than efficiency. Using the result from the set of experiments mentioned in the second step we ensured that the interaction between Match Score and Preview Length is significant (Partial F-test, t=19.145, p-value=0).

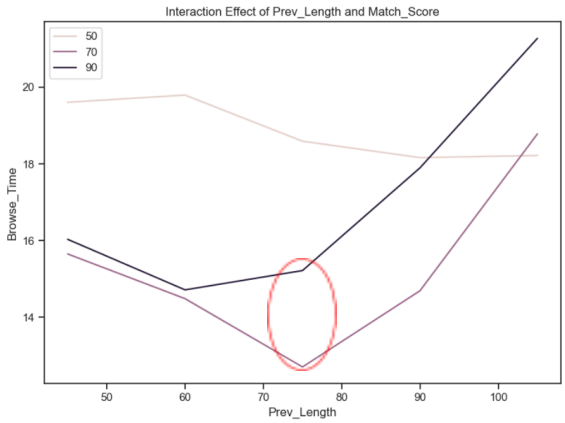
Since we established that preview type TT performs better than AC, our goal was to determine the combination of match score and preview length that results in the shortest browsing times.

The lowest average browsing time was 10.21 for Preview Length=75, Match Score=70 among the conditions mentioned in step two. In comparison, the first step experiments showed the lowest time as 16.9 for P.L.=60, M.S.=70. A hypothesis test conducted before further experiments around P.L.=75, M.S.=70 indicated the need for more exploration in this area. The test results (t=46.79, p-value=1.01e-109) confirmed a notable reduction in browsing time under these conditions, guiding our focus on enhancing these optimal settings. Histogram graph of the samples with setting P.L. = 75, M.S. = 70:

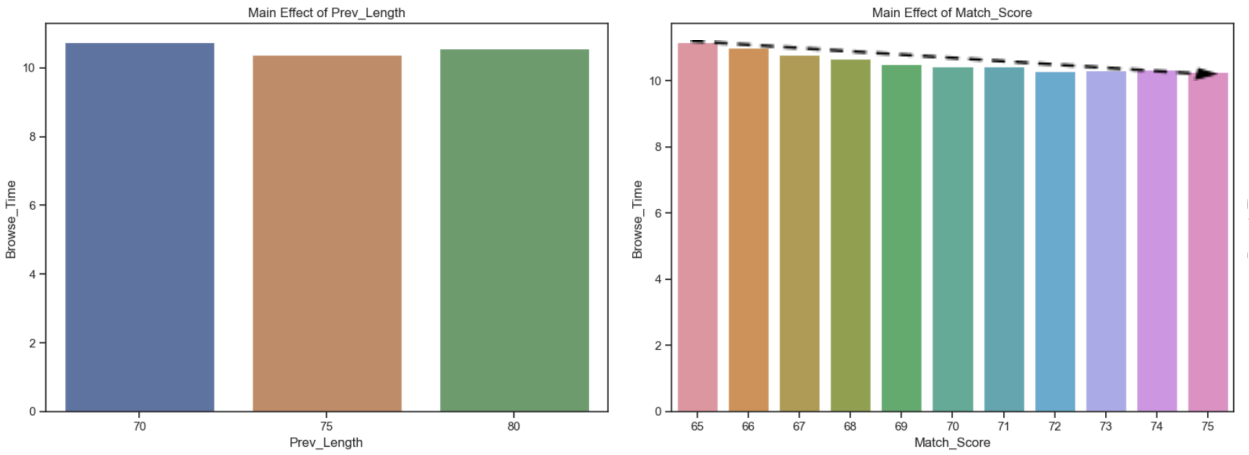


Step Four: Searching for the "Sweet Spot"

Our final objective was to accurately identify the sweet spot that consistently yields the best browsing time. Step Three provided us useful information on the nebulous region where the optimum might be, but we wanted to leverage this information, adopting a more rigorous approach to maintain efficiency in the number of experiments conducted. The interaction graph below using the data gathered from the experiments (mentioned in the second step) shows browsing times rise sharply after prev length 75 with match scores 70 and 90:



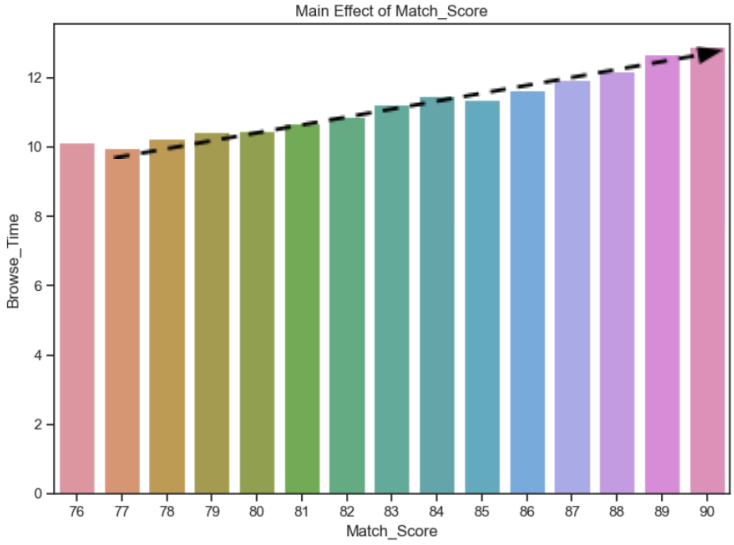
That gave us a clue to keep the preview length close to area 75 and explore more with the match scores. With that in mind, we conducted 33 trials, keeping Preview Type fixed at TT and Tile Size constant at 0.2, while systematically varying Preview Length and Match Score. Preview Length was tested at three levels: 70, 75, and 80, incrementing by 5 for each group. Within each Preview Length category, we explored Match Scores ranging from 65 to 75, increasing in single-unit steps. This methodical approach was designed to pinpoint the combination of Preview Length and Match Score that results in the most efficient Browse Time, thus determining the sweet spot within these parameters. As observed from the bar graphs below, the average browsing time reached its minimum at a Preview Length (P.L.) of 75 and continued to decrease with a slope of $B=-0.086$ ($Y=B \cdot X$) as the Match Score (M.S.) ranged from 65 to 75.



So we decided to run one more set of experiments to see which M.S. yields the minimum browsing time.

Step Five: Deciding the Boundaries of the "Sweet Spot"

Maintaining prev length 75, type TT, and tile size 0.2, we explored match scores from 75 to 90, and ran an additional set of experiments with 15 different conditions. We found a new local minimum with browsing time 9.95 at M.S. 77. As can be observed from the graph below, after a Match Score of 77, browsing times increase linearly with a slope of B=0.22 ($Y=B \cdot X$) up to a Match Score of 90.



After finding a dip at 77, in the search for the upper bound, we first conducted an F-test between Match Scores 77 and 78 to determine if their variances are equal. We failed to reject the null hypothesis that their variances were different, with an F-test result (F-statistic = 0.043, p-value = 0.836), indicating that the variances are statistically equal. Subsequently, we conducted a Student's t-test, which resulted in a t-statistic of -2.147 and a p-value of 0.033. This led us to reject the null hypothesis that their average browsing times are the same at the significance level of 0.05. Therefore, we established the upper boundary of the Match Score as 77.

Then, in the search for the lower boundary of the Match Score (M.S.), we proceeded with the same methodology until we rejected the null hypothesis. We began by comparing average browsing times starting from M.S. 76, utilizing data from the prior set of experiments for M.S. values below 76. In all these comparisons, the p-value of the F-test results was higher than 0.05, which led us to conduct Student's t-tests. Up until M.S. 71, we kept failing to reject the null hypothesis that their average browsing times were different. However, when comparing M.S. 77 with 71, we rejected the null hypothesis with an observed t-test statistic of -2.7766 and a p-value of 0.0060, concluding that the average browsing time for M.S. 77 and 71 is significantly different at the significance level of 0.01. Consequently, we defined the range for M.S. yielding the most effective browsing time as between 77 and 72 inclusive. The p-values of the 5 hypothesis tests were as follows: 0.123 for 77-76, 0.090 for 77-75, 0.246 for 77-74, 0.527 for 77-73, and 0.620 for 77-72, indicating that we fail to reject the null hypothesis at the significance level of 0.05 for those comparisons.

Conclusion

Through our experimental journey, we concluded that the most effective configuration for reducing user browsing time lies within a specific parameter range: a prev length of 75, match scores ranging from 72 to 77, using type TT, and irrespective of tile size.

The precise setting we recommend for minimizing browsing time is as follows: Tile Size set to 0.2 (default), Match Score set at 77%, Preview Length set to 75 seconds, and Preview Type set to Teaser/Trailer (TT), with a 95% confidence interval for browsing time estimated to be approximately 9.78 to 10.14 minutes.

A limitation of these findings is the nature of the data collection. There is no consideration of how different age groups or cultures interact with the features in this simulation; we may not be able to generalize our results. There could be nuisance factors affecting the results. Another limitation could be the impracticality of testing every continuous combination of factor values in real-world scenarios. This suggests exploring a range of optimal values, as indicated in Step Four, is more feasible than seeking a single definitive optimum.

There could also be non-identifiable experimental units due to cookie-based experiments. Different users might use the same device, thereby causing data leakage; the same person could be in different treatments simultaneously. This limitation can introduce noise to our data.