Executive Summary

Netflix is one of the most popular streaming platforms which provides the user with a plethora of viewing options. While the high number of viewing options provide users with a wide variety of options, it also comes with the curse of *Decision Paralysis*. This may lead a user to spend more time browsing among the options, only to be fatigued and ultimately lose interest and return without watching anything. The primary goal of this online controlled experiment is to minimize the average browsing time a Netflix user spends at the home screen of the website. An increase in browsing time may be due to several factors like suitability of the recommended content, varying interests of a user, and so on. However, for the case of this experiment, we chose to focus on four major factors which are hypothesized to influence the browsing time, i.e. Tile Size, Match Score, Preview Length, and Preview Type.

We begin by experimenting with different levels of these four factors to test which factors make a statistically significant influence on browsing time. We performed controlled experiments with these factors, starting with a 2^4 factorial test to check for the significance of each factor. This helped us eliminate Tile Size as an insignificant factor. We then performed a series of sequential factorial experiments using different conditions within our reduced search space. As a result, we concluded that the following levels produced the most optimal browsing time:

Preview Type: Teaser/Trailer, Match Score: 75, Preview Length: 75 seconds

A minimum average browsing time of 9.84 minutes with a 95% Confidence Interval of (9.62, 10.07) was obtained for this optimal condition.

Introduction

In this project, we aim to reduce the time Netflix users spend browsing the suggested content. This helps improve user engagement with the platform and addresses the issue of *decision paralysis*. We define the average browsing time as the **Metric of Interest (MoI)**. The goal is to minimize our MOI through controlled experimentation on selected design factors listed below

Factor Name	**Type**	**Description**	**Region of Operability**
Tile Size	Numeric	Ratio of a content's tile to the screen height. Aspect ratio of a tile is fixed throughout the experiment. Smaller values correspond to smaller tile sizes	0.1 to 0.5
Match Score	Numeric	Prediction of how much a user will enjoy a particular content given their past viewing history. Higher scores represent higher prediction strength	0 to 100
Preview Length	Numeric	Duration (in seconds) of the preview that is displayed in each tile when a user hovers above it	30 to 120
Preview Type	Categorical	Type of content that is displayed in the previews - trailer/teaser or actual content	TT, AC

The experiments were performed in a controlled environment with 100 responses for each experimental condition. Our initial goal was to check the importance of each factor. We performed a 2^4 factorial test with two levels for each factor sufficiently spaced apart. As part of the first step, we plotted the main effects and interaction effects of each factor. We performed partial F-tests to check for the significance of terms and pairwise t-tests to find the condition which minimizes the browsing time.

Then iterative experiments were performed to refine our search as we go along to find the optimum value of these factors for the MoI. Since we are comparing the mean browsing time, we deployed the appropriate t-tests (Student's or Welch's) based on the result of the F-test for variances. We used Bonferroni adjusted p-values to address the multiple comparison problem. We stopped this process once we could not improve the MoI further.

The next section provides a detailed description of each round of experiments and documents our reasoning for our experimental design for each round. We start by explaining the 2^k factorial experiment and its conclusion. This is followed by the algorithmic workflow of the consequent rounds. Later we list down the results of each experiment round and our conclusion on the optimal condition.

Experiments

The two central questions which motivate our series of experiments are (a) “what factors are important in minimizing our metric of interest, the average browsing time?” and (b) “what is the optimal value for these factors that we should adopt to minimize the browsing time of users?”

The response variable is the browsing time, which is the length of time a user spends browsing (as opposed to watching) Netflix. We have four design factors, tile size, match score, preview length, and preview type. The experimental units are the Netflix users who are exposed to the experimental conditions defined by these four factors. We assigned 100 Netflix users randomly for each of the conditions in the experiment.

In each of the rounds of experimentation, we looked at plots and did tests like partial F-tests, pairwise t-tests, and applied corrections to the p-values keeping in mind the multiple comparison problem. We kept the significance level to 0.05 for all the hypothesis tests.

2^4 factorial test

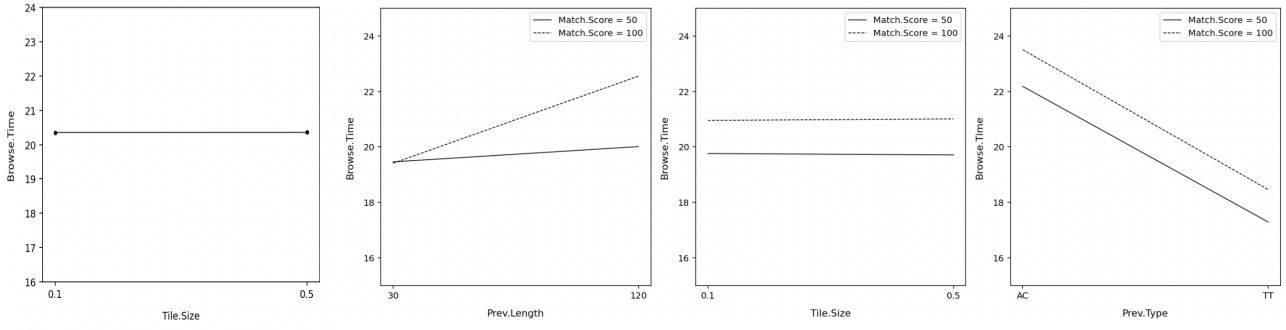
Initially, we performed a 2^4 factorial experiment with two objectives in mind - (a) to find the factors and their interactions that influence the browsing time and (b) to find the optimal experimental conditions to refine our search.

For the 2^4 factorial experiment, we take two levels for each factor and consider all the combinations. The two levels are fixed in a way that there exists a *reasonable separation.* For Match Score, we fixed the levels at [50, 100], because we expect the minimum match score for a title in the “Top Picks For. . . ” row to be 50 from a practical standpoint. For Tile Size and Preview Length, we take the entire search space at {0.1,0.5} and [30, 120] respectively. For Preview Type we choose [TT, AC].

From the factorial experiment, the most important finding was that the tile size is not relevant to minimizing the metric of interest. We did partial F-tests for checking the significance with alpha set to 0.05, starting from

- 4-factor interaction: failed to reject the null hypothesis that 4-factor interaction is not significant
- 3-factor interaction: failed to reject the null hypothesis that 3-factor interaction not significant
- 2-factor interaction: Prev Length and Match Score were significant
- Main effects: Prev Length, Prev Type, Match Score were significant

This was also supported by the main effect plots and interaction plots



Experimentation workflow: Here’s a summary of the steps we will follow in each subsequent round.

Step 1: Use **gatekeeper test** to check if at least some of all conditions are significantly different

$$H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_m = 0 \text{ vs } H_a : \mu_i \neq \mu_j \text{ for any } i \neq j$$

Proceed to Step 2 if we reject the null hypothesis. Else choose the condition with minimum Browsing time.

Step 2: Sort the experimental conditions in ascending order by their average browsing time, and select the top 3 or 4 conditions which have the least browsing time. We assume that the best condition among the current pool should be one among the top 3 or 4 closely clustered conditions. (Refer to the table below)

Conditions	Prev_Length	Prev_Type	Match_Score	Tile_Size	Mean Browsing Time
condition 1	30	TT	100	0.1	μ1
condition 2	30	TT	50	0.1	μ2
condition 3	120	TT	100	0.1	μ3
...	**...**	**...**	**...**	**...**	**...**
condition m	120	AC	50	0.1	μm

Table 1: A Demonstration of how the experiment summary statistics are arranged for testing purposes

Step 3: **Pairwise Test Procedure**: The overall goal of pairwise test design is to check whether the least browsing time condition is significantly lower than the rest of the conditions, taking

$$H_0 : \mu_1 \geq \mu_i \text{ vs } H_a : \mu_1 < \mu_i \text{ for i = 2, 3,4, ... m}$$

To check for equality of variances, we use the F test: $H_0 : \sigma_1^2 = \sigma_i^2$ vs $H_0 : \sigma_1^2 \neq \sigma_i^2$ for i = 2, 3,4, … m. Based on the number of comparisons, we applied **Bonferroni** correction, a stricter check on the p-values. All subsequent p-values are with Bonferroni corrections.

Step 4: Based on step 3, find one or more conditions with browsing time significantly lower than the rest. Refine the search region around these condition(s).

Round 1:

Following the conclusion of the insignificance of the tile size, we use existing data at 0.1 tile size for the next steps to keep n = 100. We follow the “Experimentation Workflow” described above and reject the null hypothesis of the gatekeeper test for the 8 conditions (p-value: 4.38e-07).

Now we pick the top 3 conditions and do pairwise tests,

Condition	Prev_Length	Prev_Type	Match_Score	Tile_Size	μ
1	30	TT	100	0.1	16.901
2	30	TT	50	0.1	16.951
3	120	TT	50	0.1	17.599

H_a	Reject H_0?	p-value
$\mu_1 < \mu_2$	Yes	3.26e-12
$\mu_1 < \mu_3$	Yes	2.76e-21
$\mu_2 < \mu_3$	Yes	3.45e-06

From the results we found conditions **(Preview Type, Preview Length, Match Score): (TT, 30,100)** and **(TT, 30,50)** to minimize the browsing time. Furthermore, for the levels of Prev Type factor, AC and TT, we conclude that TT has a lower average browsing time based on the values for the two groups. We assume this would be the case at least in the vicinity of our initial conditions. For subsequent rounds, we keep TT as the level for Prev Type. Now we refine the search space keeping preview length biased towards 30 and explore levels for match score in Round 2.

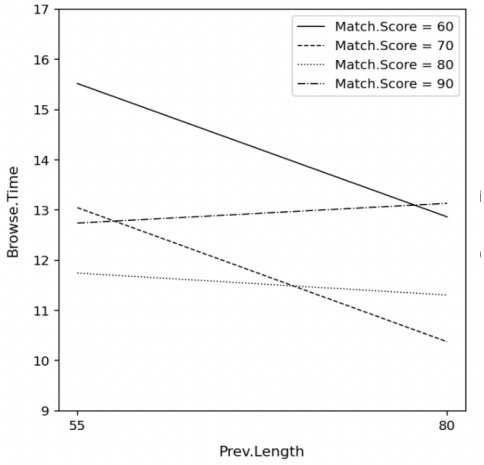
Round 2:

Levels chosen for the next round are summarized below

- Match Score: [60,70,80,90]
- Preview Length: [55,80] seconds
- Tile Size: Default Value (0.2) and Preview Type: [TT] (Fixed for all consecutive rounds)

This helps us explore more intermediate levels for Match score - as the MoI for the levels chosen in the prior round were almost identical. We anticipate the optimal point in the region. As we are more confident that the optimal region lies towards the lower end of the Preview Length scale, we choose two additional levels skewed towards 30 seconds.

From the interaction plot, we see that for lower values of match score (60 and 70) the browsing time reduces as the preview length increases. This we hypothesize could be attributed to the fact that if the match score is less (i.e. a user doesn't know much about the title), showing longer previews helps them decide to watch the title, thus reducing the browse time.



We follow the "Experimentation Workflow" established earlier for this round. From the gatekeeper test, we can reject the null hypothesis, i.e. the mean browsing time of all conditions are significantly different. We proceed forward to perform the pairwise tests.

Best conditions and Pairwise Test results:

Condition	Prev_Length	Match_Score	μ
1	80	70	16.901
2	80	80	16.951
3	55	80	17.599

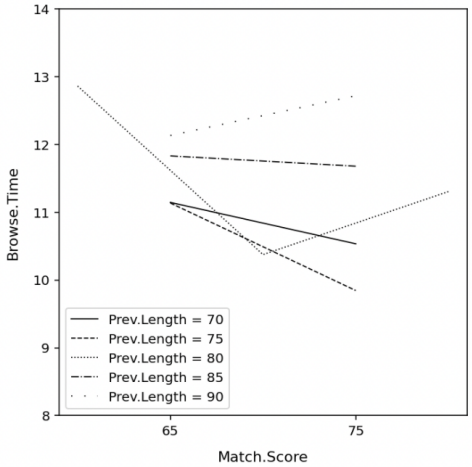
H_a	p-value	Reject H_0?
$\mu_1 < \mu_2$	1.95e-11	Yes
$\mu_1 < \mu_3$	1.65e-20	Yes
$\mu_2 < \mu_3$	3.04e-45	Yes

We conclude that **[Prev Length: 80, Match Score: 70]** gives the minimum browsing time.

Round 3:

Levels chosen for this round are further probing the region around the optimal point from the last step. However, we are reducing the distance between the new levels to be chosen.

Match Score: [65,75], one point on either side of the optimum from the last round
Preview Length: [70, 75, 85, 95], two points on either side of the optimum from the last round



The data from these 8 experimental conditions was augmented with previous conditions Preview Length = 80 and match score: [60, 70,

80] to also include the top two conditions from the previous round. From the interaction plot, we see that Prev Length of 75 and Match Score of 75 give us the lowest average browsing time.

Following the “Experimental Workflow”, we can reject the null hypothesis of the gatekeeper test again.

The best experiment conditions and pairwise test results are shown below:

Condition	Prev_Length	Match_Score	μ
1	75	75	9.84
2	80	70	10.38
3	70	75	10.53

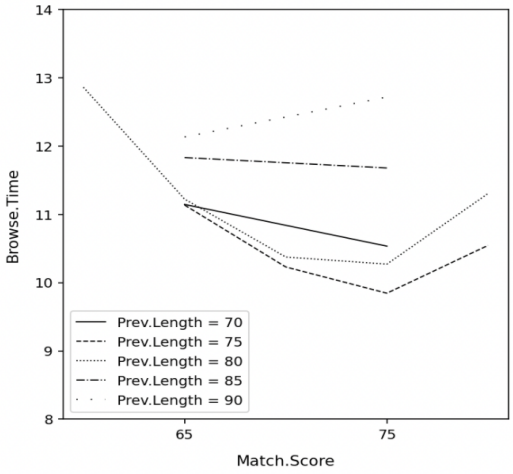
H_a	p-value	Reject H_0?
$\mu_1 < \mu_2$	3.28e-04	Yes
$\mu_1 < \mu_3$	1.6e-05	Yes
$\mu_2 < \mu_3$	2.07e-05	Yes

We conclude that **[Prev Length: 75, Match Score: 75]** gives the minimum browsing time.

Round 4:

Now we are confident that the minimum does lie around the region we have explored so far, so we want to fill in any missing points around it. Therefore we chose the following levels missing from our existing test data.

- Prev Length : 75 seconds; Match score: [70, 80]
- Prev Length : 80 seconds; Match score: [65, 75]



The data from these 4 experimental conditions were added to data from the last round. Looking at the interaction plots now gives us confidence that Preview Length of 75 seconds and Match Score of 75 does indeed look like the minimum.

Following the experimentation workflow, the results obtained are shown below:

Condition	Prev_Length	Match_Score	μ
1	75	75	9.84
2	75	70	10.23
3	80	75	10.27

H_a	p-value	Reject H_0?
$\mu_1 < \mu_2$	0.022	Yes
$\mu_1 < \mu_3$	0.015	Yes
$\mu_2 < \mu_3$	1.000	No

For completion, we also checked the values for Match Score:[74, 76] at Preview Length of 75 secs and found that the browsing time significantly increased ($\mu \sim 10$). To test if Preview Type = TT is better, we also checked the average browsing time near our optimum for Preview Type = AC and we see a significant increase in browsing time ($\mu \sim 15$). As a consequence of the above tests, we can conclude that the levels which produced the most optimal browsing time are: **Preview Type: Teaser/Trailer, Match Score: 75, Preview Length: 75 seconds.** A minimum average browsing time of 9.84 minutes with a 95% Confidence interval of (9.62, 10.07) was obtained for this optimal condition.

Conclusion

Based on our testing, we concluded that three of the four factors, i.e Preview Type, Preview Length and Match Score significantly influence the browsing time. We could not establish any statistical significance of changing levels in the Tile Size factor. Hence, we excluded it in our analysis after the first 2^4 factorial screening round. After four additional iterative rounds of sequential experiments to shrink our search space, we found the lowest average browsing time of 9.84 minutes with a 95% Confidence interval of (9.62, 10.07) corresponded to the experimental condition:

Preview Type: Trailer Type, **Preview Length:** 75 seconds, **match score:** 75

We need to be cognizant of a few limitations in our approach:

1. Nuisance factors that are not controlled - like the time of the day during which the experiment was performed. It could be argued that experiments performed during office hours could attract a different kind of viewers than those performed during other times during the day.
2. We did not perform power analysis in the experiment and worked with 100 units in each condition. This was allowed as per the instructions but in the real world, we need to perform a power analysis to arrive at the minimum number of units in each condition.
3. We could perform a polynomial fit using the central composite design method to find the optimal levels at a more granular level.