

Netflix Homepage Browsing Time Optimization

Group Members: Emma Rainer, Zhiyi Ren, Dominnic Chant

Final Project - MSDS 629

Executive Summary

When a user goes onto the Netflix homepage, they see different rows of selections of movies or shows. We wanted to focus in particular on the “Top Picks for...” row that outputs suggestions curated for the given user. However, a problem for the user may arise called decision paralysis, where a user is overloaded with too many choices, making it hard for the user to make a decision. This problem may eventually lead the user to leave the service without watching anything. We want to minimize browsing time so that a user can find their movie or show quickly, lessening the chance of decision paralysis. We found the optimal condition that will minimize browsing time through a series of data gathering and experimenting on certain conditions. We reached the lowest browsing time at 10.16 minutes by setting Tile Size to be 0.2, Preview Type to be teaser/trailer, Match Score to be 75%, and Preview Length to be 75 seconds.

Introduction

When Netflix members browse the homepage and hover over the display artwork of a movie, the recommendation algorithm will automatically play the preview clip of the movie. However, due to the large number of movies displayed on the homepage, users cannot choose which movie to watch in the short time when viewing the preview. Therefore, in order to improve the user experience on Netflix, our team wants to reduce the browsing time of users by running the simulator and performing factorial experiments. We want to get a result that users can minimize browsing time by four factors: Match Score, Tile Size, Preview Length and Preview Type¹. After we successfully understand whether user browsing time is associated with the main effects, interaction effects, or both and test for significance, we will find a better way to help users find the types of stories they were looking for within their Netflix experience and gain more positive feedback from end-users. We followed the QPDAC strategy to derive the optimized solution.

Initially, we raised the question of how to reduce members' browsing time and developed an experimental plan, then collected data through a simulator and made five rounds of experiments, including statistical hypothesis analysis and exploring whether interaction effects were significant. Our experiments are based on a factorial experimental design. This method divides each design factor into two or more levels to explore the effects between the factors and find the best combination by comparing different combinations of factors to obtain the lowest average browsing time.

We first set the reasonable range for each design factor, conducted 16 experimental conditions, and explored whether interaction effects are significant by fitting a full linear regression model and using a partial F test to compare a reduced model. The results indicated that Tile Size is not a significant factor of browsing time, and Preview Length and Match Score have a significant interaction effect. We also found that Preview Type has a main effect on browsing time only, and the teaser/trailer condition results in a lower browsing time than the actual content condition. Therefore, we fixed the Preview Type condition as TT in the subsequent experiments.

Our second experiment explored the three significant design factors; we further narrowed the range based on the interaction effect plot. Since the last experiment showed a significant change in average browsing time when the Match Score was 100, we lowered the upper limit to 75. In this step, we conclude that Match Score is a significant predictor.

For the third experiment, we conducted four conditions and learned that a Preview Length of 75 seconds resulted in a lower average browsing rate at a Match Score of 70 and 80. The browsing rate was lower for a Match Score of 70.

In the fourth experiment, we fixed the range of matching scores and controlled the Preview Length between 70 and 80. However, the level of Preview Length makes a significant difference to the browsing time. Therefore, we used Preview Length equal to 75 for the next experiment.

In the fifth experiment, we combined the data from the above multiple experiments to check the t-test. We fixed the Match Score as 75 and compared the Preview Length to 70 and 75. Our p-value is 0.253, indicating there is no significant difference in average browsing time between a Match Score of 70 and 75 when Preview Length is 75 and the other design factors are default. At this point, we selected the observed average browsing time as the lowest factor. Finally, we got the optimal solution is: [Match Score=75, Preview Length=75, Tile Size =0.2, Preview Type =TT]

¹**Tile Size:** The ratio of a tile's height to overall screen height

Match Score: Predict how much users enjoy watching the featured content (in percentages) based on the viewing history

Preview Length: The duration (in seconds) of a movie/show preview.

Preview Type: The type of previews played automatically

Experiments

We are interested in whether any of the alterable Netflix homepage factors that comprise this row influence average browsing time among viewers and, if so, what combination of the factors minimizes the average browsing time. As such, we experimented with the size of the preview tiles, the length of the preview shown, the type of preview that is autoplayed (trailer or actual content), and the match score prediction of how much a viewer will enjoy a show or movie. The factors of interest and their regions of operability are summarized in the table below:

Factor	Region of Operability	Default Value
Tile Size	[0.1 ,0.5]	0.2
Match Score	[0, 100]	95
Preview Length	[30, 120]	75
Preview Type	{TT,AC}	TT

Experiment I

The objective of the first experiment was to determine the relationship between the “Top Picks For …” row factors and browsing time. We utilized a 2^4 experimental design where our metric of interest was average browsing time and the corresponding response variable was the total browsing time spent during a customer’s visit to the homepage. The experimental units were the Netflix home page visitors. The design factors and the levels we chose to experiment with are summarized below:

Design Factors	Levels
Tile Size	{0.1, 0.5}
Match Score	{50, 100}
Preview Length	{30, 120}
Preview Type	{TT,AC}

We chose 0.1 and 0.5 as the levels to test for Tile Size since the region of operability is narrow and we wanted to test two levels that were sufficiently different from each other. For Match Score, we picked 50% and 100% as our starting levels. We reasoned that since larger values indicate a higher likelihood of enjoyment, then it does not make sense to suggest content with smaller values ($< 50\%$) with which we expect the customer to have a low level of enjoyment. For Preview Length, we did not have a strong inclination as to whether Netflix thought a customer might prefer a longer or shorter preview since the default fell exactly in the middle. Thus, we chose the extremes of the region of operability. Finally, for Preview Type we tested both available levels.

In total, 16 experimental conditions were created from the combinations of our factor levels. Each homepage configuration was shown to 100 customers and their average browsing time was found. Of these 16 configurations, we were interested in which resulted in the lowest average browsing time.

To test which interaction effects are significant, we fit a full linear regression model to the data:

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4 + \beta_8 x_2 x_3 + \\ \beta_9 x_2 x_4 + \beta_{10} x_3 x_4 + \beta_{11} x_1 x_2 x_3 + \beta_{12} x_1 x_2 x_4 + \beta_{13} x_1 x_3 x_4 + \beta_{14} x_2 x_3 x_4 + \beta_{15} x_1 x_2 x_3 x_4$$

where

- $x_1 = 1$ when Preview Type is AC, and 0 otherwise
- $x_2 = 1$ when Preview Length is 120, and 0 otherwise
- $x_3 = 1$ when Match Score is 100, and 0 otherwise
- $x_4 = 1$ when Tile Size is 0.5, and 0 otherwise

The t-tests of $H_0 : \beta_j = 0, j = 1, 2, \dots, 15$ show that, at a significance level of 0.05, only Preview Type, Preview Length, Match Score and the Preview Length:Match Score two-way interaction are significant indicators of browsing time. We confirm this with an ANOVA test comparing the full model to the reduced model - $Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_8 x_2 x_3$. We test:

H_0 : Reduced model fits the data as well as the full model

H_A : The reduced model does not fit the data as well as the full model

The resulting p-value is 0.87. At a 0.05 significance level we reject the null hypothesis and conclude the reduced model explains the variance in browsing time as well as the full model and, therefore, confirm that the terms we removed were not statistically significant indicators of browsing time.

Based on this experiment, we determine that Tile Size is not a significant predictor of browsing time, so in further experiments it can be set to the default. We also conclude that Preview Type has a main effect on browsing time only, so from looking at the main effect plot (Figure 1a) we determine that the teaser/trailer condition results in a lower browsing time than the actual content condition.

Preview Length and Match Score have a significant interaction effect, so we look at the interaction effect plot to get insight into how the factors interact. We can see that Match Score has a stronger effect of Preview Length when Preview Length is 120.

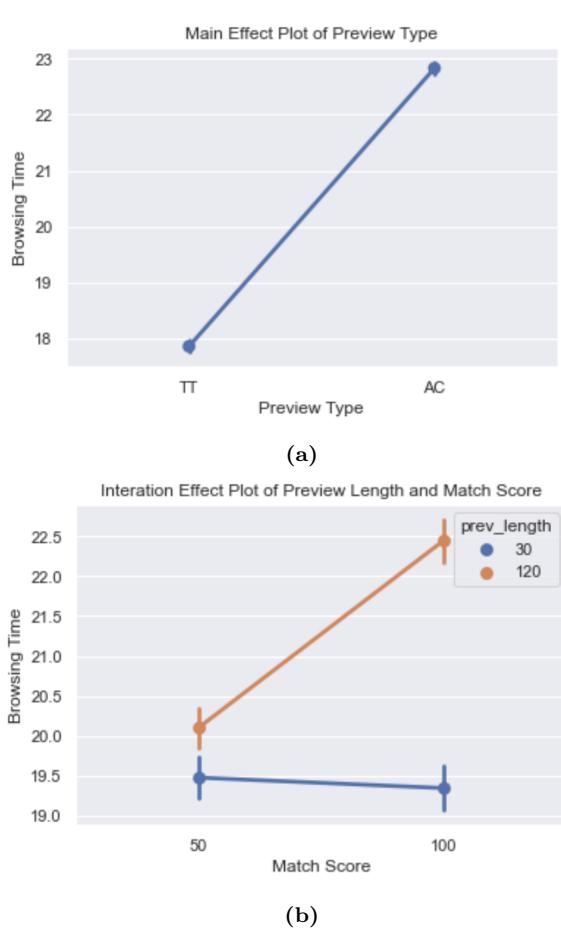


Figure 1

Experiment II

For our second experiment, we built upon the findings in Experiment I to answer: what is the optimal configuration of the three significant predictors (Match Score, Preview Type and Preview Length) to minimize browsing time?

We utilized a 2^2 experimental design. Again, our metric of interest was average browsing time and the corresponding response variable was the total browsing time spent during a customer's visit to the homepage. For Experiment II, we let Tile Size and Preview Type equal their defaults (0.2 and TT). We chose to test Match Score levels of 50 and 75 because the average browsing time was relatively low for both Preview Lengths at a score of 50, but there was a big variation in average browsing time when Match Score was 100, so we lowered the upper level. In addition, we chose to test Preview Length at levels 45 and 90 since it was not conclusive whether either extreme was preferable. Values closer to the median might be more insightful. The design factors and the levels we chose to experiment with are summarized below:

Design Factors	Levels
Match Score	{50, 75}
Preview Length	{45, 90}

Based on the interaction effect plot, it is apparent that the Match Score of 75 results in a lower average browsing time for a preview length of both 45 and 90. Fitting a linear regression model to the data, the t-test for Match Score ($H_0 : \beta_{MatchScore} = 0$) results in a p-value of 0.00. At a 0.05 significance level, we determine that Match Score is a significant predictor, and thus the mean browsing time is significantly different for a Match Score of 50 and 75. The average browsing time is lowest for a Match Score of 75 and Preview Length of 45 (12.37).

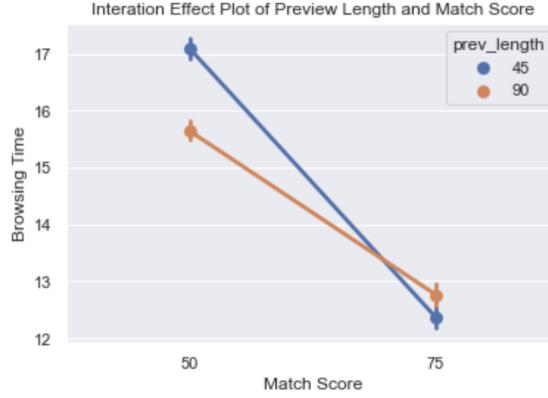


Figure 2

Experiment III

For our third experiment, we wanted to further optimize the configuration of factor levels. Building off of Experiment II, we utilized a 2^2 experimental design with the same metric of interest and response variable. We decided to test Match Score at levels of 70 and 80- a 5% change in either direction of our previously observed optimal Match Score. We also decided to change preview length to 60 and 75 in order to observe more values. The design is summarized below:

Factors	Levels
Match Score	{70, 80}
Preview Length	{60, 75}

As visualized in the interaction effect plot below, from this experiment we learned that a Preview Length of 75 seconds resulted in a lower average browsing rate at both a Match Score of 70 and a Match Score of 80. The browsing rate was lower for a Match Score of 70.

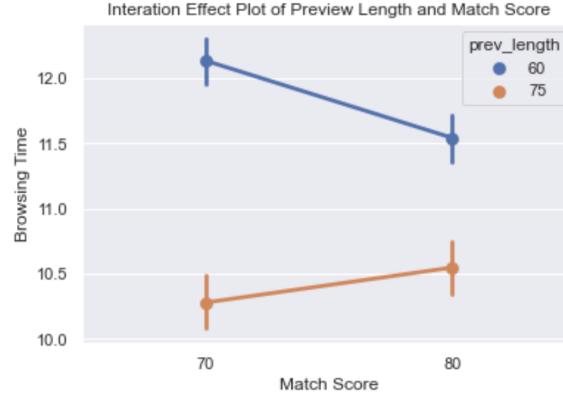


Figure 3

Experiment IV

For our fourth experiment, we conducted another 2^2 experimental design to try to find the configuration of factors that minimize the mean browsing time. We tested a Preview Length of 70 and a Preview Length of 80- a 5% change in either direction of our previously observed optimal Preview Length. We tested Match Score at the same levels as in Experiment III:

Factors	Levels
Match Score	{70, 80}
Preview Length	{70, 80}

The optimal configuration from these four conditions was a Preview Length of 80 and a Match Score of 70 which resulted in a mean browsing time of 10.50 minutes. However, at a Preview Length of 70 and a Match Score of 70, the mean browsing time observed was 10.57 minutes. As such, we decided on a Preview Length of 75 since it fell in between the two levels we experimented with.

Experiment V

Since we had resources available, we chose to conduct a 2^1 experiment to see if a Match Score of 70 or 75 resulted in a smaller average browsing time when Preview Length is fixed at 75 minutes. We found that there was no significant difference. We fit a linear regression model and the t-test for Match Score ($H_0 : \beta_{MatchScore} = 0$) resulted in a p-value of 0.253. This tells us that there is no significant difference in average browsing time between a Match Score of 70 and 75 when Preview Length is 75 and the other factors are default.

Experimental Results

We select the configuration of factors at which the observed average browsing time is lowest. We observed the lowest average browsing time at a Match Score of 75, a Preview Length of 75, a Tile Size of 0.2 and a Preview Type of Teaser\Trailer (TT). This value is 10.16.

Using a t-score with $n - 1 = 99$ degrees of freedom (1.98), we calculate a 95 percent confidence interval for our estimate of the minimal average browsing time to be (9.952, 10.376).

Conclusion

After a total of five experiments, we concluded that the most optimal condition to minimize browsing time down to 10.16 minutes was to set Tile Size to be 0.2, Preview Type to be teaser/trailer, Match Score to be 75%, and Preview Length to be 75 seconds. Doing so should lead to a browsing time of about 10.16 minutes. During our experimentations, we found that there was no statistically significant difference between setting Match Score within a 70%-75% range, given that the other conditions stayed the same. Since there was no significant difference, we decided to go with a 75% Match Score to match the user's preferences better.