MSDS 692
Experiments in Data Science

Final Project

Team 15
Yu-Chuan Chiu
Abhradeep Mukherjee
Daniel Tinoco
Ruirui Zhang

Executive Summary

Netflix users often struggle to determine what show they would like to watch, and as the time they spend browsing increases, the more likely they are to logout of Netflix in favor of another experience or task. To remedy this, we conducted a series of experiments, including factorial tests and pairwise t-tests, to identify the significant factors and the most optimal experimental set of conditions for a user's homepage that minimize the average time spent by a user browsing their homepage for a show to watch. Our findings indicate that a combination of tile size of 0.2, preview length of 70 seconds, preview type as teaser trailer, and match score of 80% give us an expected minimum average browsing time between 8.7 minutes and 12.61 minutes with a 95% level of confidence.

Introduction

Netflix is one of the world's largest online television platforms with a huge volume of content in the form of movies, series, documentaries, etc. As a user, we can log in to Netflix with a few easy clicks, and after a few seconds, what is commonly referred to as "Zombie Scrolling Syndrome" kicks in, where we endlessly browse content until we ultimately refrain from choosing a show to watch at all.

Our goal is to reduce the browsing time of a user by optimizing certain factors or aspects of the homepage. The factors that we are focusing on in this project are -
- Tile Size: The size of the preview button of a movie/ series.
- Preview Length: The length of the preview of the movie/series shown to a user to draw their attention.
- Preview Type: The type of preview, i.e., a teaser/trailer or an actual scene.
- Match Score: This tells us how close the match is for a specific user.

On finding the right combination of the above factors we aim to assist the user in determining their choice faster, which in turn will diminish the likelihood of the user logging off or disengaging from Netflix.

This report consists of the entire journey of this project and how we found the optimal set of combinations of the factors that lead to the expected minimum average browsing time. In the remaining portion of the report the reader will be able to understand:
- The questions we were trying to answer
- The plan/approach we took to answer these questions
- The data collection process
- The analysis of the data
- The conclusion based on our findings

The Experiments

Given the QPDAC outline described above, we are interested in answering the following questions:
1. Can we determine tile size to be insignificant in the very beginning giving us more space to experiment with the other factors?
2. What is the optimal combination of tile size, match score, preview length, preview type that could minimize the browsing time?

From these questions, we planned out two phases of observation gathering to determine the answers to both questions. First, we would collect data using four combinations of the four factors, with 2 combinations of default values with tile size of 0.2 and 0.4, respectively, and another two combinations where the other factors were set at equal values distinct from the default ones and a tile size of 0.2 and 0.4, respectively. From there, we would conduct pairwise t-tests to determine the significance of tile size on the browsing time, both as a singular factor and via interaction with the other three factors. If we found that tile size was significant, then we would carry it forward into the second phase of observation gathering; if not, we could safely disregard the factor and set tile size to the default 0.2.

From there, we would construct a factorial experiment using all the other factors as well as tile size if it was found to be significant, followed by paired t-tests for determining the best levels for each factor from the factorial experiments. Aside from tile size, each design factor's levels are shown below:

| Factor	Levels
Match Score	{70, 80, 85, 90}
Preview Length	{30, 60, 90, 120}
Preview Type	{'TT', 'AC'}

From the first phase we were able to rule out tile size as a significant factor; thus, in order to search for an optimal browsing time, we would obtain data with 32 (4x4x2) conditions. We would then identify significant factors and interactions through an iterative process of creating "full" models with all factors and their interactions being included and then removing components with insufficient evidence to show they were affecting browsing time as "reduced" models, continuously comparing and iterating until we could reach a final model where all the factors and interactions included were significant in affecting browsing time. Below plot shows that tile size has no significant impact on browsing time .
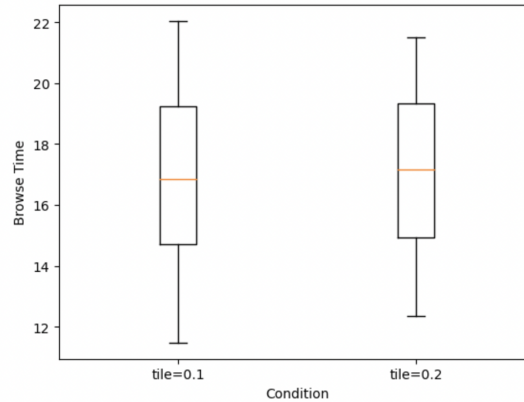
Through our iterative process, we determine the following:

- First iteration: insufficient evidence to determine the interaction between Match Score, Preview Length, and Preview Type as significant with 95% confidence
- Second Iteration: insufficient evidence to determine the interaction between Match Score and Preview Type, Preview Type and Preview Length as significant with 95% confidence
- Third and Fourth Iterations: sufficient evidence to determine that the remaining individual factors (Match Score, Preview Length, Preview Type) as well as the interaction between Match Score and Preview Length are significant in affecting browsing time with 95% confidence

From there, we conducted the pairwise t-tests to identify factors that influenced browsing time in a lower direction. Below, the boxplot shown helps illustrate that the browsing time with Teaser Trailer preview type has a significantly lower average browsing time.
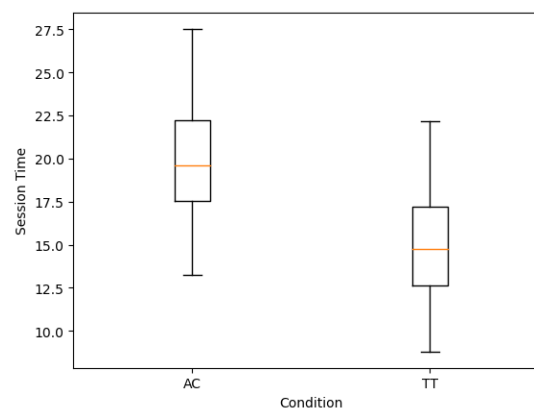
After confirming with a pairwise t-test, we found sufficient evidence that the Teaser Trailer preview type was significant in reducing the browsing time compared to Actual Content; thus, since this factor had no other significant interactions with the other two, we fixed Teaser Trailer as our optimal parameter for Preview Type.

The factorial test shows that the interaction between match score and preview length is significant. We further group the data using match score and preview length and evaluate whether there is a condition with a significant shortest browsing time.
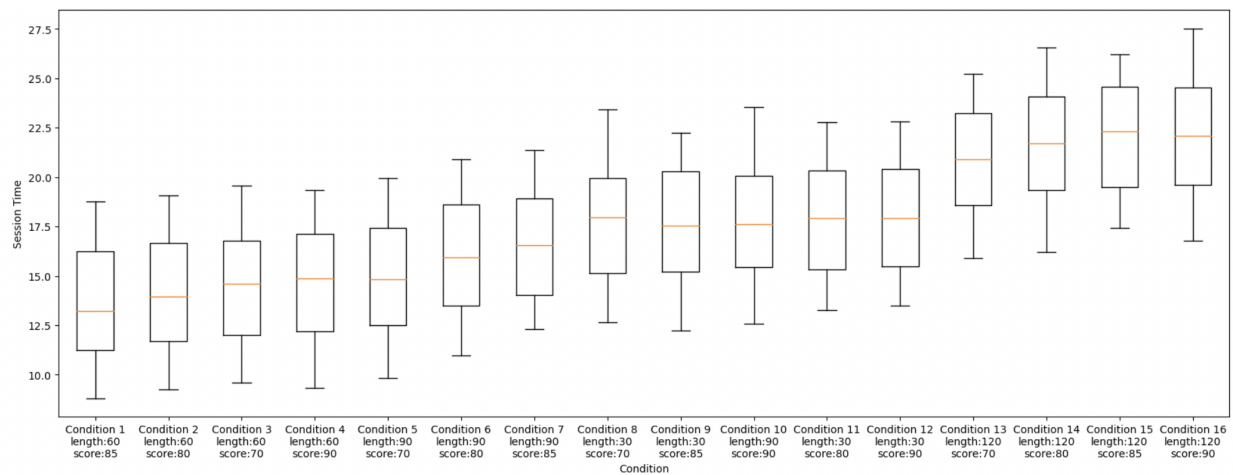

- We first test if there is a significant difference between the average browsing time for condition 1 (preview length: 60 and match score: 85) and condition 2 (preview length: 60 and match score: 80). Through f-test, we find out that the two tested conditions have the same variance.
- We thus follow up with a two-sided Student's t-test to test the average browsing time. Since the p-value of the Student's t-test is greater than 0.05, we cannot reject the null hypothesis and conclude that the average browsing time for condition 1 and condition 2 are not significantly different.
- We proceed the same process for condition 1 and condition 3. However, this time, we yield a Student's t-test p-value less than 0.05. This suggests that the average browsing time for condition 1 and condition 3 are significantly different.
- We finally conduct a one-sided Student's t-test between condition 1 and condition 3. The result shows that the average browsing time for condition 1 is significantly smaller than condition 3. These tests indicate that both condition 1 (match score: 85, preview length: 60, tile size: 0.2, preview type: TT) and condition 2 (match score: 80, preview length: 60, tile size: 0.2, preview type: TT) can give us the smallest browse time with no significant difference. We would prefer condition 2 over condition 1 since a lower match score indicates more movies can be displayed.

With the results obtained above, we have a seemingly optimal condition set of the following:
- Tile Size: 0.2
- Preview Length: 60 seconds
- Match Score: 80%
- Preview Type: Teaser Trailer

However, we wished to further refine or justify our results using Response Surface Methodology Below are the contours for Actual Content and Teaser Trailer, respectively, with tile size fixed at 0.2.

Predicted Average Browsing Time for Teaser Trailer — Predicted Average Browsing Time for Actual Content
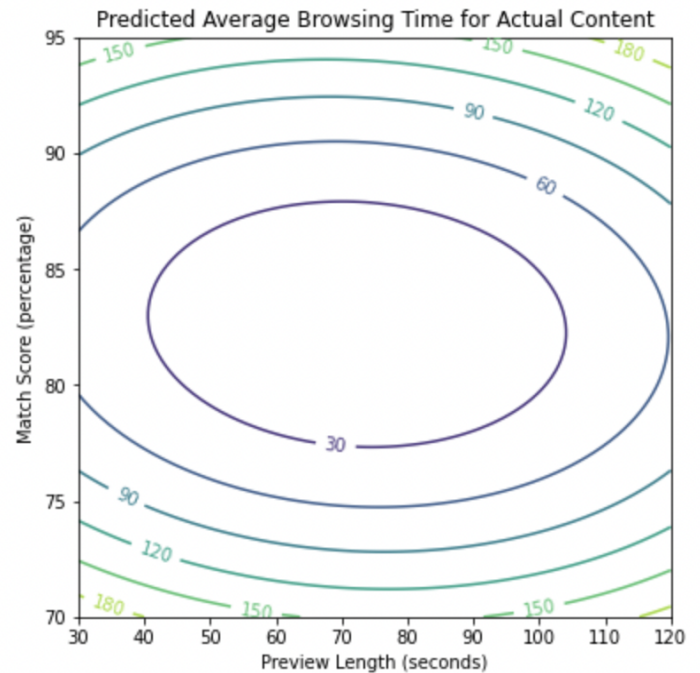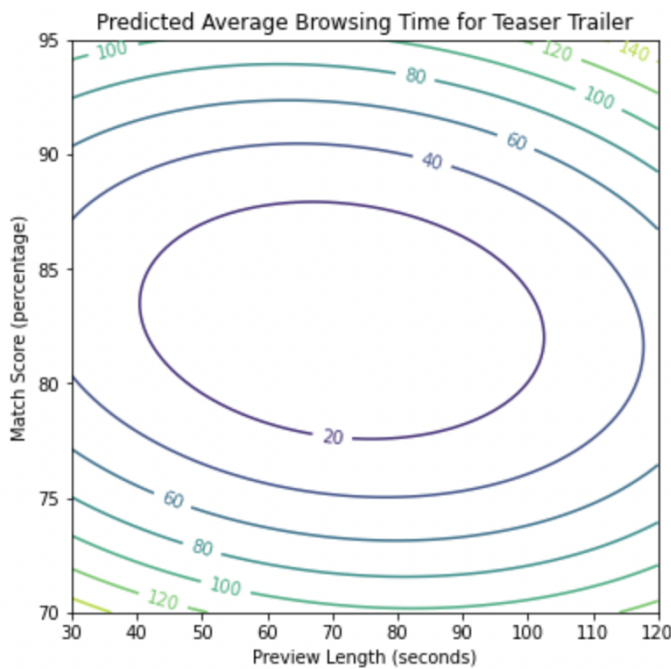
Since we have shown that Teaser Trailer is more significant than Actual Content, we ignore the contour for AC and consider the result for TT from the contours(keeping tile size as 0.2) which are:

- Preview Length: 72
- Match Score: 83

We observe that the match score value is almost similar to what we have calculated before but there is a significant difference in preview length. Therefore to further improve on our results we generated data with two more conditions which are:

Condition 1:
- Tile Size: 0.2
- Preview Length: 65 seconds
- Preview Type: Teaser Trailer
- Match Score: 80%

Condition 2:
- Tile Size: 0.2
- Preview Length: 70 seconds
- Preview Type: Teaser Trailer
- Match Score: 80%

From the above, we determined that there is a significant difference between the average browsing time between 60, 65, and 70 seconds of preview length, and found that 70 seconds of preview length obtains the optimal interval for the expected minimum average browsing time which is in-line with the results that we obtained from the contours of response surface methodology.

Conclusion

Our experiments show that by setting tile size at 0.2, preview length at 70 seconds, preview type as teaser trailer, and match score at 80%, we obtain the interval for the expected minimum average browsing time to be between 8.7 minutes and 12.61 minutes with a 95% level of confidence.

Limitations

First, we must clarify that our findings do not declare the *expected* minimum average browsing time interval to be the *actual* minimum average browsing time interval.

Furthermore, we must acknowledge that due to the limited resources at our disposal, our methods for creating experimental conditions to test are not exhaustive enough to test every possible combination of the 4 factors, meaning that there are literal gaps in our experimentation that might have revealed better results. We were also limited to obtaining 100 observations for each set of conditions, perhaps preventing us from obtaining an estimate closer to the true value of average browsing time for those conditions. Lastly, we specifically chose to hone in on a match score above 70% purely for heuristic reasons. Thus, it is entirely possible that there exists a very low match score, or just a lower match score, that leads to better results.