MSDS 628 – Experiments in Data Science

Final Project Report

01/20/2023

Matt Wheeler | Lanting Su | Joy Hueng | Bharadwaj Allu

Executive Summary

We performed A/B experimental analysis by varying four design features of the Netflix website; namely tile size, match score, preview length, and preview type. Our goal was to find the optimum feature values that minimized the average time users spent browsing the website. By performing a 2k factorial analysis, response surface optimization, and pairwise t-test grid search, we conclude that tile size does not affect user browsing time and that the optimum values for match score, preview length, and preview type are 75%, 74s, and teaser trailer preview respectively.

Introduction

Netflix is a multi-billion dollar internet streaming company that hosts an extensive library of tv-shows, movies, and documentaries. A core concern within Netflix is that such variety can cause choice paralysis amongst its users. The more time a user spent browsing, and trying to decide what to watch, the more likely they are to ultimately lose interest and choose not to watch anything at all. To address this issue, we will use A/B experimentation to optimize Netflix's website to reduce the time spent browsing and consequently prevent users from being dissatisfied and leaving.

To do this, we will experiment with different configurations of the following website features; tile size, match score, preview length, and preview type. We will perform a 2k factorial analysis to identify which features significantly affect the user browsing time, and subsequently use partial F-test analysis to statistically verify the results. Following this we will find the approximate optimum feature values using second order response surface optimization methods. Finally we will looking to improve the accuracy of these optimum feature value predictions by performing a localized pairwise t-test grid search.

Experimental Analysis

The general approach to A/B experimental analysis is to define a metric of interest that you desire to optimize, establish a set of variable design factors that are believed to influence this metric of interest, vary those factors for different users, and record the metric of interest in each case. By doing so, you can understand what, if any, effect each design factor has on the metric of interest, as well as their optimum values. In the context of this investigation, we wanted to minimize the time Netflix users spent browsing for something to watch. There were four website features that we were able to alter that we established as the design factors. These features were:

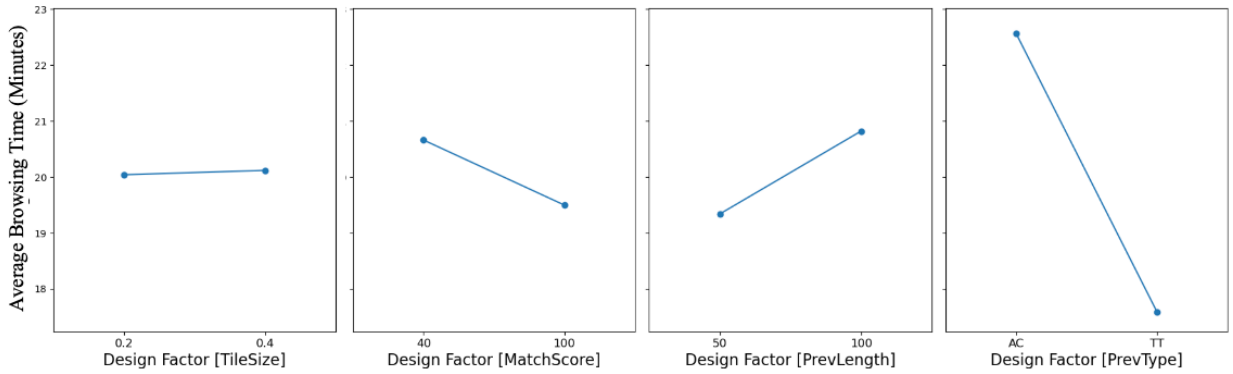
- **Tile Size**: The ratio of the tile height to the overall screen height
- **Match Score**: The probability, expressed as a percentage, that a user would enjoy a show or movie based on their viewing history
- **Preview Length**: The duration, in seconds, of the show or movie's preview
- **Preview Type**: The type of preview shown to the users

These design factors had a set range of values that they could take. Tile size, match score, and preview length were all continuous values that ranged from 0.1 – 0.5, 0 – 100%, and 30 – 120s respectively, while the preview type was either actual content or a short teaser trailer.

To understand the effect each design factor had on browsing time, we first performed a 2k factorial analysis. This involved selecting an upper and lower value of each design factor. These values are displayed in the table below. While the tile size and preview length values were evenly spaced between their extremes, the match score lower and upper values were selected as 40% and 100% respectively because we believed the optimum value would lie towards to upper end of the range.

Design Factor	Lower Value	Upper Value
Tile Size	0.2	0.4
Match Score	40	100
Preview length	50	100
Preview Type	Actual Content	Tease Trailer

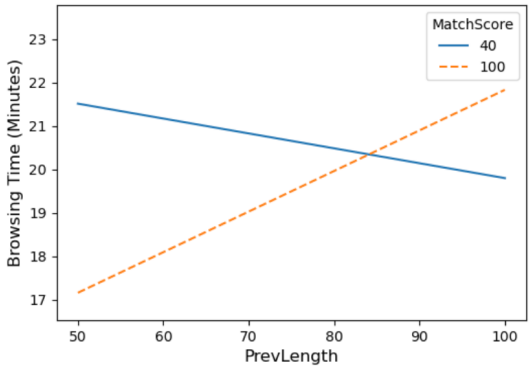
These upper and lower values resulted in 16 unique combinations of design factor values. We assigned each set of unique conditions to 100 Netflix users and recorded the average browsing time of each user. We used this data to determine the impact each of the design factors had on user browsing time. The recorded data was grouped by the unique values of each design factor and the mean browsing time was plotted, as seen in the main effect plots below.



From these plots we can see that tile size doesn't appear to impact browsing time, whilst increased match score, decreased preview length, and teaser trailer previews all appeared to reduce user browsing time. We were able to statistically validate these conclusions by performing several OLS regression partial F-tests. These tests compared two models: the first model, referred to as the full model, encoded all design factors as categorical variables and included all feature interaction terms; the second model did the same but excluded one of the factors. This process was repeated until all design factors had been excluded from the second model. These partial F-tests were able to test the null hypotheses that the design factor missing from the second model was not significant.

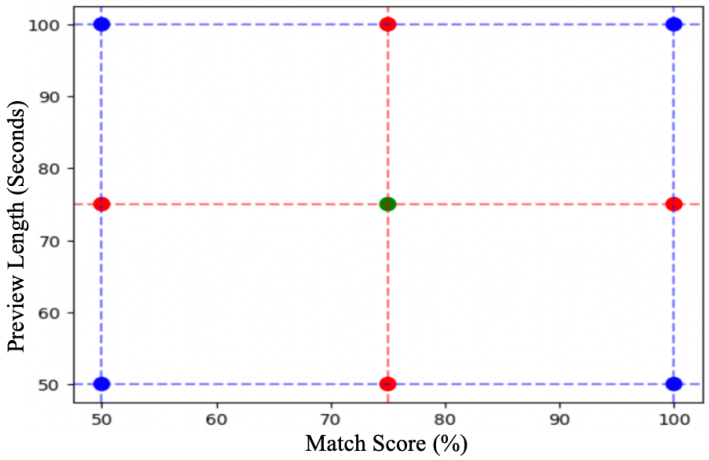
The p-values for match size, preview length, and preview type were all computed to be equal to zero. As a result, we can say there is sufficient evidence to reject the null hypotheses and conclude that all three design factors impact the browsing time. Conversely, tile size was computed to have a partial F-test p-value of 0.205. Since this is greater than our 5% significance level, we can say there is insufficient evidence to reject the null hypothesis and conclude that tile size does not impact the user browsing time.

The regression model summary, for the above defined full model, indicated that the only significant feature interaction term was between preview length and match score. The interaction plot below shows that when match score equaled 100, longer preview lengths increased browsing time. Contrarily, when match score equaled 40, longer preview lengths decreased browsing time.



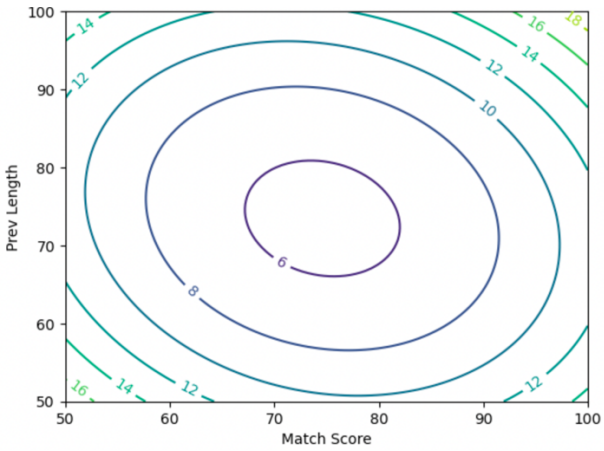
Given that there were no significant interactions between preview type and the other design factors, we were able to perform a student's t-test to determine the optimum preview type. To do this, we tested the null hypothesis that the mean browsing time for users who were shown a teaser trailer was greater than or equal to the mean browsing time for users who were shown actual content. The corresponding p-value was computed to be 4.36×10^{-305}. Since this value is much less than our 5% significance level, we can say there is sufficient evidence to reject the null hypothesis and conclude that the optimum preview type is a teaser trailer.

Once we had eliminated tile size from our analysis and established that teaser trailer was the optimum preview type, our next objective was to determine the optimum values for match score and preview length. To do this, we performed a two-factor central composite design to simulate a second-order response surface that would approximate the relationship between average user browsing time and the match score and preview length design factors. To fit this model, we assigned 200 Netflix users to each of the nine match score and preview value combinations represented by the dots in the plot below. For each condition set we ensuring that the preview type was set to teaser trailer and that the tile size was set to the default value of 0.2.



To improve the performance of this approximation, we reduced the upper and lower value range for match score from 40 – 100% to 50 – 100%. We did this because when we ranked the mean browsing time for each condition in our original 2k factorial analysis, we saw that the feature combinations with the lowest mean browsing time mostly had match scores equal to 100%. This indicated, that the true match score optimum lay at the higher end of the original range. Also, by increasing the number of users assigned to each condition set from 100 to 200, we hoped to improve the robustness of the model.

We prepared the collected data by applying relevant transformations from natural to coded units. We then fit a second-order linear regression model, the results of which suggested that all the main interaction effects are significant in the model. Below is the contour graph, in natural units, obtained with a clean convex shape, which confirmed that the condition ranges we chose were appropriate. We observe from this plot that the range for the optimum match score is between 70% and 80%, with the approximate optimum estimated to be 74.6%. Similarly, the range for the optimum preview length is between the 70s and 80s, with approximated optimum estimated to be 73.5s. A slightly less optimal, but practically necessary, step was to round these values to the nearest integer. Hence, we chose 74s and 75% for approximate optimum or preview length and match score respectively.



To further evaluate that our approximated optimum conditions are accurate, we performed a grid search of points localized around the optimum. For each grid search point, we performed a student's t-test to test the hypothesis that the mean browsing time for the approximate optimum conditions was greater than or equal to the mean browsing time for the grid search point. We selected the following grid search points to test:

[1] Match Score: 80, Preview length: 79
[2] Match Score: 70, Preview length: 79
[3] Match Score: 80, Preview length: 69
[4] Match Score: 70, Preview length: 69

The p-values associated with these hypothesis tests are displayed in the table below. Each of the stated p-values are less than our specified 5% significance level, meaning we can say there is sufficient evidence to reject each of the null hypotheses and conclude that mean browsing time for the approximate optimum is less than the mean browsing time for each of the grid search points surrounding the optimum.

Condition 1	**Condition 2**	**p-value***
Preview Length: 74, Match Score: 75	Preview Length: 79, Match Score: 70	0.0458
Preview Length: 74, Match Score: 75	Preview Length: 69, Match Score: 70	3.41×10^{-4}
Preview Length: 74, Match Score: 75	Preview Length: 69, Match Score: 80	7.15×10^{-6}
Preview Length: 74, Match Score: 75	Preview Length: 79, Match Score: 80	1.15×10^{-11}

* The p-value corresponding to the student's t-test of the null hypothesis that the average browsing time from condition 1 is greater than the average browsing time from condition 2.

Conclusion

By performing an initial 2k factorial analysis, modelling an approximation browsing time response surface, and conducting a pairwise t-test grid search, we can conclude two things: firstly, that the tile size ratio has no impact on the user browsing time, and secondly, that the optimum conditions

for preview length, match score, and preview type are 75s, 74% and teaser trailer previews respectively.

There were a few limitations that we faced during this experiment that may have affected our results, the first being the limited number of condition combinations that we could check. Given more time and resources, we could have performed a more extensive pairwise grid search with the aim of confirming, or establishing a new, optimum condition set. Secondly, being limited to only assigning 100 Netflix users to each condition combination may have affected the robustness of our results. In reality we would have access to a much larger pool of users.

Further experimentation could be conducted to conclude if these limitations did in fact affect our results, but this would require significantly more resources than is currently available.