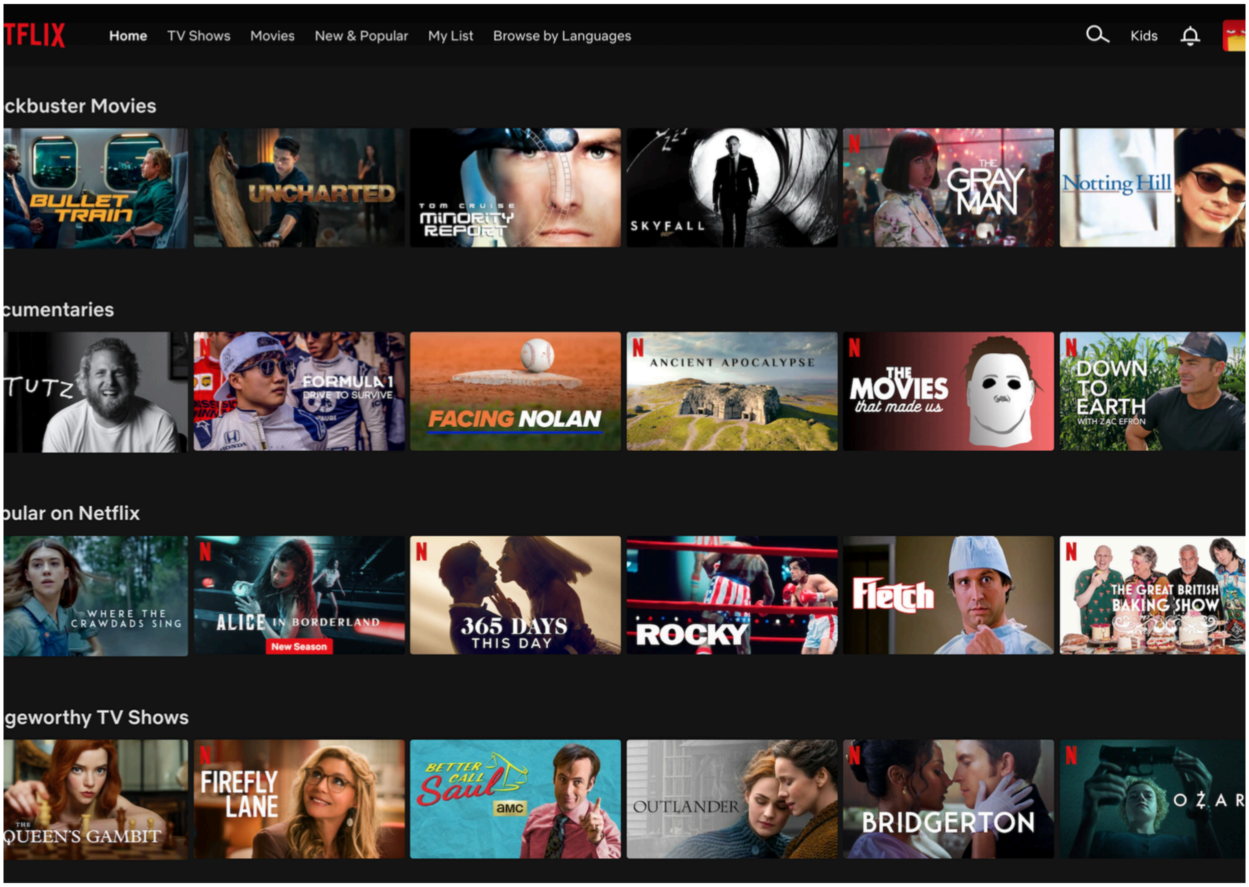


NETFLIX A/B EXPERIMENT

MSDS 629

Claire Zhou **Nick Miller**

Sangjun Han **Irene Garcia Montoya**

EXECUTIVE SUMMARY

This report details the experimental process done to identify the best combination of four design factors of the Netflix homepage, aiming to minimize users' browsing time (BT). To do so, a series of 2^K factorial experiments followed by a Central Composite Design experiment were conducted. With the new found configuration, the browsing time will be minimized to between 9.93 and 10.37 seconds with a 95% level of confidence.

INTRODUCTION

Netflix has set out to reduce the average browsing time by helping their users make a decision through a curated “Top Picks For. . . ” row of the homepage. In order to find the best design of this curated row, we have designed a series of A/B experiments with the objective of finding the optimal combination of the following design factors in order to reduce the average browsing time (response variable):

➢ *Tile Size*: size ratio of the show’s row tile.
➢ *Match Scor*e: percentage of the estimated likelihood that the user will enjoy the show/movie.
➢ *Preview Length*: duration of the shown preview.
➢ *Preview Type:* show either the teaser/trailer (TT) or a snippet of actual content (AC).

Our experimental approach encompassed a comprehensive series of statistical designs, starting with a series of 2^K factorial experiments followed by a Central Composite Design (CCD) experiment. The initial 2^K factorial experiment served the purpose of assessing the individual and combined influence of the design factors. Thanks to it, we identified Tile Size as non-significant, which allowed us to exclude it from the rest of the experiments, as well as determining that ‘TT’ was the best level for Preview Type going forward. As for Match Score and Preview Length, we further deciphered their optimal values by conducting three additional rounds of 2^2 factorial experiments.

Once we thought we were close enough to their optimal values, we employed the CCD methodology to pinpoint the specific factors’ optimal levels and used our model to predict the new average BT.

Finally, a conclusive experiment was conducted using the determined optimal conditions to validate our findings as well as calculating the confidence interval of the response variable under our new configuration.

THE EXPERIMENTS

The data

For each of our experimentation rounds, we recorded the browsing times of 100 users for each of the conditions we designed. Each condition was designed within the region of operability of each factor, taking into account some restrictions:

Design variable	Region of Operability	Restrictions
Tile Size	[0.1,0.5]	
Match Score (%)	Integer ∈ [0,100]	Must be an integer
Preview Length (seconds)	Float ∈ [30, 120]	Can only be changed in 5-second increments
Preview Type	{AC, TT}	

2^K Factorial experimentation

To start, we wanted to investigate which of the factors have influence over the response variable and get a sense of which levels yield better results. To do so, we employed a systematic approach using 2 to the K factorial experiments, conducted over four distinct rounds. As stated before, these experiments aimed to identify the area of the hyperspace where the optimum value of the response variable may be located.

First Round of Experimentation

The initial phase involved all four factors, so we conducted a 2^4 factorial experiment. For the purpose of the experiment, we needed to decide on two distinct levels for each factor. We chose:

Design variable	Low Level	High Level
Tile Size (TS)	0.1	0.5
Match Score (MS)	25	75
Preview Length (PL)	50	100
Preview Type (PT)	TT	AC

The outcomes of the experiment showed that, under a 0.05% significance level:

- ➢ MS, PL and PT have significant main effects on the BT
- ➢ There exists a significant two-way interaction between MS and PL.
- ➢ By a slim margin, we observe a three-way interaction between MS, TS and PT. However this was later disputed by a partial F-test where we tested a reduced model (without the interaction) against the full model. The F-test concluded that there is no evidence to support that the full model is better than the reduced model, and therefore the inclusion of the three-way interaction is not justified.

These conclusions allowed us to conclude that going forward the value to TS is irrelevant as it does not impact the BT in any way. On the other hand, by observing the main effect plot of PT (Figure 1.0) and the interaction effect of MS with PL (Figure 1.1) we also concluded the following:

- ➢ Of the two levels of PT, TT yields the best BT results, which means that going forward we will set this factor on TT.
- ➢ Out of the tested combinations of MS and PL, a lower level of PL (50) coupled with a higher level of MS (75), results in a lower average BT.

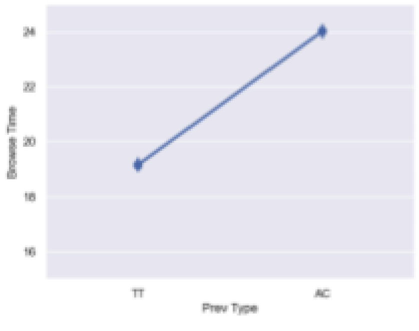


Figure 1.0

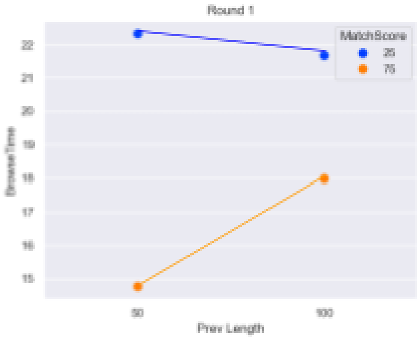


Figure 1.1

Second round of experiments

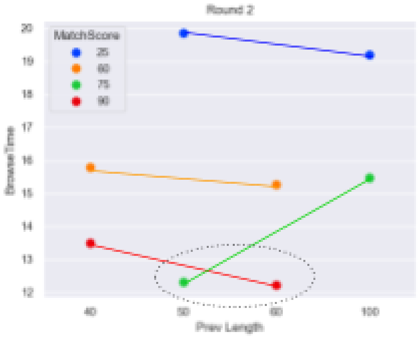


Figure 1.2

In the second round, the experimental focus narrowed down. Since we have fixed TS to its default value and PT to TT, we only need to continue experimenting around the best performing combination of MS and PL. We adjusted the ranges to 40 and 60 for PL, and 60 and 90 for MS, spanning four experiments in total (Figure 1.2). We also included the data collected during the first round that pertains to PT equal to TT. This allowed us to have more information. The combined results seemed to point to the existence of an optimal BT value in the area around PL = [50,60] and MS = [75, 90].

Third and Fourth Round of Experiments

Our third and fourth rounds of experiment served to search deeper into this area of interest (AoI). We started by setting PL to the levels of 45, 55 and 65 and MS at 75 and 95. From these experiments we were surprised by finding an even better BT result on the extreme left of the AoI, at PL = 65 and MS = 75 (Figure 1.3). Before proceeding to the Central Composite Design (CCD) phase, we decided to explore PL values exceeding 65, hypothesizing the existence of more optimal points. Consequently, we set PL at 65 and 75, and MS at 70 and 80. This final round discovered even better results of the average BrowseTime (very close to 10 seconds) (Figure 1.4).

Feeling we were close enough to the optimal, we decided to proceed to the CCD phase, with MS = 70, PL = 75, PT=TT, TS = 0.2 as our best configuration so far.

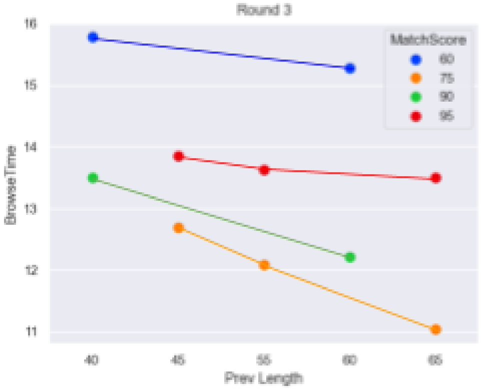


Figure 1.3

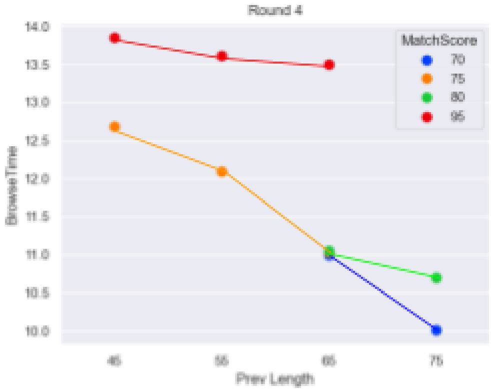


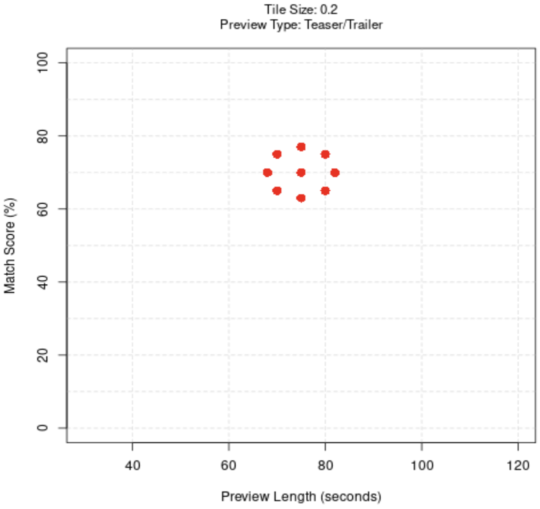
Figure 1.4

CCD (Central Composite Design)

The final step of our experimental journey was to use a central composite design to find the condition that minimizes average browsing time. The success of the CCD in finding an optimum value is dependent on selecting values that are near the optimum which is why the previous rounds of experimentation and data collection were critical.

Due to our previous findings that Tile size was not a significant factor and that Teaser/Trailer (TT) was a significant main effect with no interactions, we were able to utilize a CCD design with K' = 2 and *a* = 1.4. The two design factors which would be included in the second order model were Preview Length and Match Score. This design required us to simulate another sample of 9 conditions from these two factors. The values of these conditions were:

- Two level factorial conditions with values based on the prior rounds of experiments (4 Total)
    - PrevLength = [70, 80]
    - MatchScore = [65,75]
- Axial conditions determined by the value of *a* (4 Total)
    - PrevLength = [68,82]
    - MatchScore = [63,77]
- A single center point (1 Total)
    - PrevLength = [75]
    - MatchScore = [70]



A visualization of the conditions is shown to the right.

Once we simulated data from these conditions, we had to convert the factor values back from natural units to coded units in order to use them in the linear regression model. The coded units took on the values of -1 and +1 in order to represent the low and high values of each factor, while a value of 1.4 was used for *a,* and a value of 0 was used for the center point.

We next fit the regression model shown below and identified that the p-values associated with both of the quadratic effects in the model were approximately zero. Thus, we could reject both of the null hypothesis and conclude that the quadratic effects were significant and there was some optimum that could be identified by our model:

$$\eta = \beta_0 + \beta_1 PrevLength + \beta_2 MatchScore + \beta_{12} PrevLength * MatchScore + \beta_{11} PrevLength^2 + \beta_{22} MatchScore^2$$

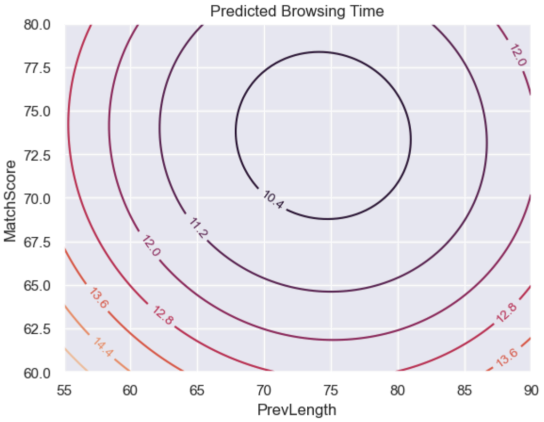
We next calculated the stationary point and converted it to natural units. We rounded these natural units in order to fit the project parameters and make them adhere to the business use case. Using these values for the stationary point, we were then able to calculate the minimum browsing time predicted by the regression model.

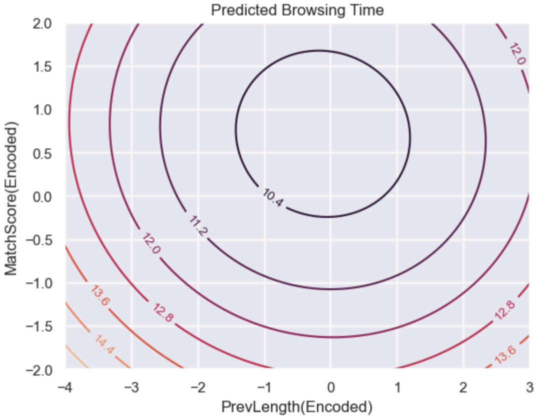
Stationary Point in Coded Units: PrevLength = -.01125, MatchScore= .7145

Stationary Point in Natural Units (Not Rounded): PrevLength = 74.437, MatchScore = 73.5726

Stationary Point in Natural Units (Rounded): PrevLength = 75, MatchScore = 74

Predicted Minimized Browsing Time: 10.0790 seconds



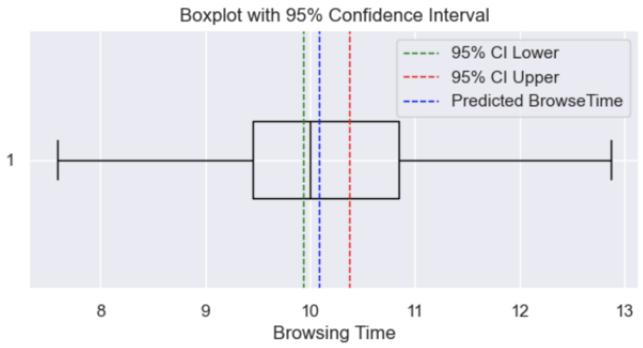


Final Verification

From the final CCD results we see that the BrowseTime is minimized by PrevLength= 74.437 and MatchScore= 73.572. Rounding them to valid values, we set the following combination of factors as the optimal:

- Match Score = 74, Preview Length = 75, Preview Type = TT and Tile Size = 0.2 (default value)

With this combination, our model estimates that the Browsing Time will be around 10.08 seconds. Finally, we confirm the validity of these results with one last experimentation round and from its results we calculate that, with the factors' levels stated above, the Browsing Time of users will be between 9.936 and 10.373 seconds with a 95% level of confidence.



CONCLUSION

To the Netflix Design Team, we recommend configuring the homepage's curated row so it shows a Match Score of 74%, a Teaser/Trailer preview that's 75 seconds long to minimize the average browsing time of users. The size of the tiles is not influential and can be configured with its current default value.