NETFLIX HOMEPAGE “TOP PICKS” OPTIMIZATION

Anish Mukherjee, Mengting Xu, Zihao Ren

Executive Summary

The project's goal is to prevent decision paralysis, which is achieved by minimizing the user's browsing time. We experimented with four factors (Tile Size, Match Score, Preview Length, Preview Type).

To find the optimal conditions, we decided to run experiments to help us. The process is as follows:

1. We ran a 2 ^ 4 Factorial experiment to determine which factors significantly impact the users browsing time.
2. We ran several experiments on these screened factors and found the optimum factor value.
3. We utilized hypothesis testing methods to see whether there was a statistically significant result.

After running several experiments, our optimal condition is:{Tile Size: not significant effect, Match Score:75, Preview Length:75, Preview Type: TT} with an average user browsing time of 10.11 minutes.

Introduction

Research shows that people may feel it becomes more complex and take longer to make decisions when they face many options, so do Netflix users; this is called decision paralysis. Netflix wants to avoid decision paralysis by helping users choose what to watch quickly. Minimizing users' browsing time on the homepage can avoid users losing interest, thus maintaining user retention and increasing conversion rate. Based on this motivation, we ran a series of experiments with different conditions to see which would minimize our metric of interest: the average user browsing time.

We will dig into 4 different factors: Tile Size, Match Score, Preview Length, and Preview Type, to find out the best combination of these factors.

Factor	**Definition**	**Data Type**	**Region of Operability**
Tile Size	The ratio of a tile's height to the overall screen height	float	[0.1,0.5]
Match Score	A prediction of how much you will enjoy watching the show or movie	integer	[0,100]
Preview Length	The duration (in seconds) of a show or movie's preview	integer	[30,120]
Preview Type	The type of preview that is auto-played	categorical	{TT, AC}

In the first phase, our goal is to screen out which factors are essential to users' browsing time and which are not. The factorial approach essentially means investigating all the combinations of levels of multiple factors are experimented then finding the optimal one. We used a particular case of the factorial experiment with just two extreme levels for each factor which can minimize the number of experimental conditions required to investigate multiple factors so that could improve our experiment efficiency. We will explain how we did this in detail in the experiments part. To improve our experiment efficiency, we also selected two groups of sample data with the two different Preview Type while all the other factor levels were equal. We chose the F-test and two-sample Student's T-test to find the better one among the two Preview Type levels.

Since we found the Tile Size did not contribute to our final goal and fixed the level of Preview Type in the first phase, we would focus on the other 2 factors: Match Score and Preview Length in phase 2. We tried several combinations of factor levels to determine which combination of levels is optimal. We also chose the F-test and T-test mentioned above to select the optimum location. Since each step deals with the multiple comparison problem, the Bonferroni Correction was also utilized by us to control the Family-Wise Error Rate(FWER), which stands for the probability of committing a Type I Error in any of the M hypothesis tests.

Experiments

Phase 1: Factor screening.

Our first objective is to drop features that are unimportant to browsing time. We used the 2^4 factorial experiments to guide our study.

For the 2^4 factorial experiments, we took two levels for four factors, which gave us 16 conditions in total. The two levels for each factor were chosen to be at extreme ends of the region of operability. This would allow adequate opportunity for the factor to express itself. For Match Score, we decided the levels 5 and 95, for Preview Length, we chose the levels 35 and 115, for Tile Size, we chose 0.15 and 0.45, and Preview Type as TT and AC; these were our experimental conditions.

After gathering the data, we fitted a linear regression model containing all the factors, which we called our full model. Then we dropped the factor that we wanted to test the importance of, and this would be our reduced model. We compared the two models using a partial F-test: if the p-value coming from the F test is less than 0.05, we deem the factor important in predicting browsing time; otherwise, we would exclude it from subsequent analysis.

Experiment Conditions: Tile Size: {0.15, 0.45}, Match Score: {5, 95}, Preview Length: {35, 115}, Preview Type: {TT, AC}

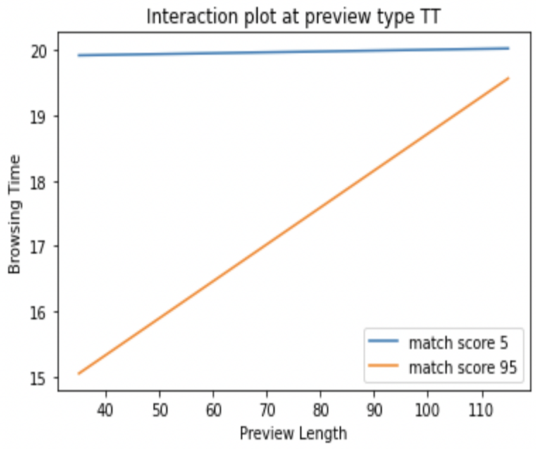
Test the significance of each factor:
Full model: Browse_time ~ C(Prev_Length) * C(Match_Score) * C(Tile_Size) * C(Prev_Type)
Reduced model0: Browse_time ~ C(Prev_Length) + C(Match_Score) + C(Tile_Size) + C(Prev_Type)
Reduced model1: Browse_time ~ C(Match_Score) * C(Tile_Size) * C(Prev_Type)
Reduced model2: Browse_time ~ C(Prev_Length) * C(Tile_Size) * C(Prev_Type)
Reduced model3: Browse_time ~ C(Match_Score) * C(Prev_Length) * C(Prev_Type)
Reduced model4: Browse_time ~ C(Match_Score) * C(Prev_Length) * C(Tile_Size)

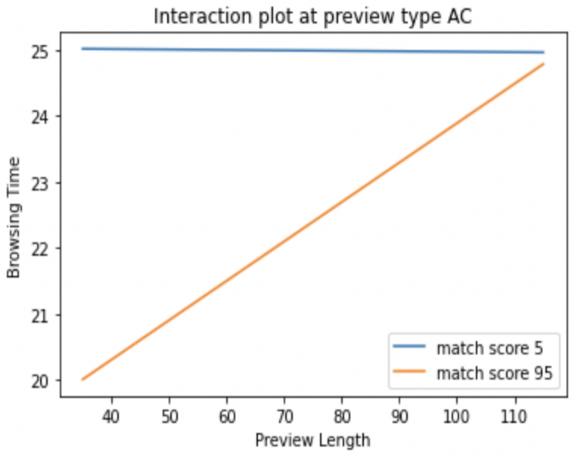
Reduced Model	RM0	RM1	RM2	RM3	RM4
P-value	5.537243e-277	0.0	0.0	0.995648	0.0

Result:

- We firstly found the interaction effects were significant. And we found that Tile Size was insignificant to predicting browsing time and hence we dropped it.
- With a p-value close to zero and t statistic value of 11.869, we found that the interaction of Preview Length with Match Score was significant.

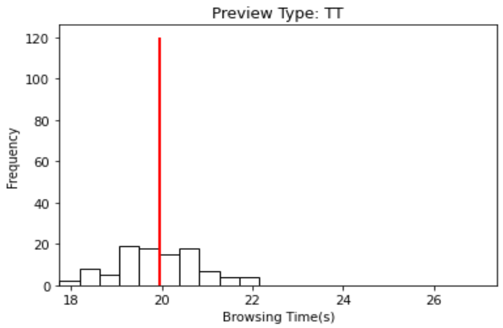
Fixing Preview Type:

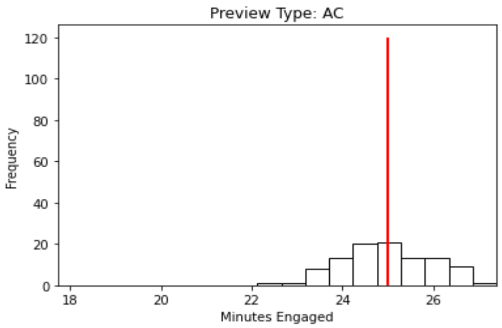




We plotted the browsing time against Preview Length for different Match Scores to improve our experiment efficiency using the sample data we collected above. The two plots represented the observations taken at Preview Type teaser trailer and Preview Type actual content. As we can see, both graphs look roughly similar, except the graph for the Preview Type teaser trailer seems to be shifted by 5 minutes downward relative to Preview Type AC.

To make this statement mathematically rigorous, we tried to test the hypothesis that browsing time for users watching teaser trailers is less than users watching actual content. But before that, we also used F-test for an equal variance to determine whether we should use Welch's t-test or Student's t-test. We also implemented this F-test and T-test pipeline to help us make the right decision in the following part.





The p-value from F-test is 1.4905. The t statistic from the Student's t-test came out to be 37.16 with a p-value of 1.6 e-91, indicating that users watching the teaser trailer had significantly less browsing times than users watching the actual content.

Result:

- This led us to fix one of the minimum conditions: Preview Type with **teaser trailer.**

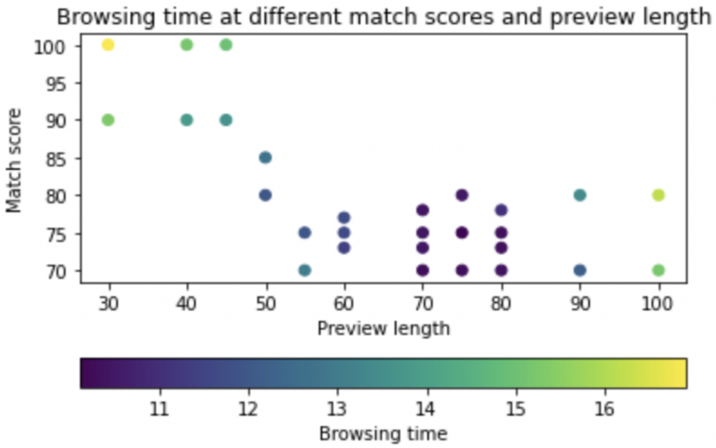
Phase 2: Narrow down the scope.

We demonstrated earlier that tile size was irrelevant in predicting browsing time. Additionally, the preview type was fixed as teaser trailer based on Welch's t-test. Therefore, the only variables we wanted to vary were the preview length and match score. Since the interaction between these two factors was significant, we adjusted them in tandem to determine the local minimum browsing time**.**

Based on the 2 ^ 4 factorial experiment results, our minimum browsing time occurred at Preview Length 35, Match Score 95, and PreviewType TT, which we verified through a multiple comparisons Welch t-test. We then decided to explore the neighborhood of this condition. The experimental conditions we took were as follows.
(preview length, match score): (30, 90), (40, 90), (45,90), (30,100), (40,100), (45,100)

The conditions (40, 90) and (45,90) had the least observed browsing time. We compared the average browsing time with the other conditions surveyed using a one-sided hypothesis test and suitable Bonferroni corrections to establish this statistically. We rejected the null hypothesis that the browsing times were equal in favor of the alternative that the browsing time for match score 90 and preview length 40/45 were significantly less than the nearby points.

The direction of this new minimum was along with decreasing match score and increasing preview length. So for the next round of explorations, we increased preview length and decreased match score and compared the minimum browsing time with the existing minima using Welch's t-test or Student's t-test based on the F-test for equal variance. We would update the minima if the new conditions had a browsing time that was statistically less than the previous minima and proceed in the same direction; otherwise, we would leave the minima point unchanged and change direction. In the end, we would assign our condition to be the local minima if the experimental conditions within the neighborhood had a statistically longer browsing time. These steps led to a minimum browsing time at Match Score 75 and Preview Length 75. Below is a plot of all our explorations.



Verification of minima

After obtaining the minimum browsing time at Match Score 75 and Preview Length 75, we decided to collect more data around the vicinity of this condition. We test for the hypothesis that the browsing time at our minimum condition is statistically less than the points within the neighborhood.

The neighborhood of interest was Preview Length at {70, 75, 80} and Match Score between 70 and 80.

Again we performed a Welch t-test to test this hypothesis.

However, this time, we could not get a statistically significant result against the hypothesis that the conditions within the neighborhood have browsing time greater than the minimum condition point. Therefore we conclude that the minima lie in the range defined by Preview Length: {70 to 80} and Match Score: {70 to 80}. However, since our observed browsing time was lowest at Match Score and Preview Length 75, we reported the condition as our minimum location.

Conclusion

Finally, Our optimum final pick is:{Tile Size: not significant effect, Match Score:75, Preview Length:75, Preview Type: TT}. The average user browsing time's 95% confidence interval is between (9.92, 10.31) minutes.
But our conclusion still has limitations, and limitations lie in that we cannot get a statistically significant difference in the final four conditions in this experiment journey. For Match Score, we can safely conclude that its optimum level lies somewhere between 70 and 80; for Preview Length, the best level lies between 70 and 80 seconds. However, we picked and recommended the experimental conditions that show the smallest sample mean user browsing time. In reality, the tradeoff of collecting more data near the minimum would lead to spending resources without a significant decrease in user browsing time. Our experiments improved the metric of interest from 15 minutes if we used the default setting to somewhere near 10 minutes, showing a 33% improvement. We feel this would improve Netflix users' experience and help the platform improve customer retention. Finally, we believe the experiment's theme of letting the data speak journey can benefit our future professional life as data scientists.