Experimental Design and Analysis for User Response Optimization

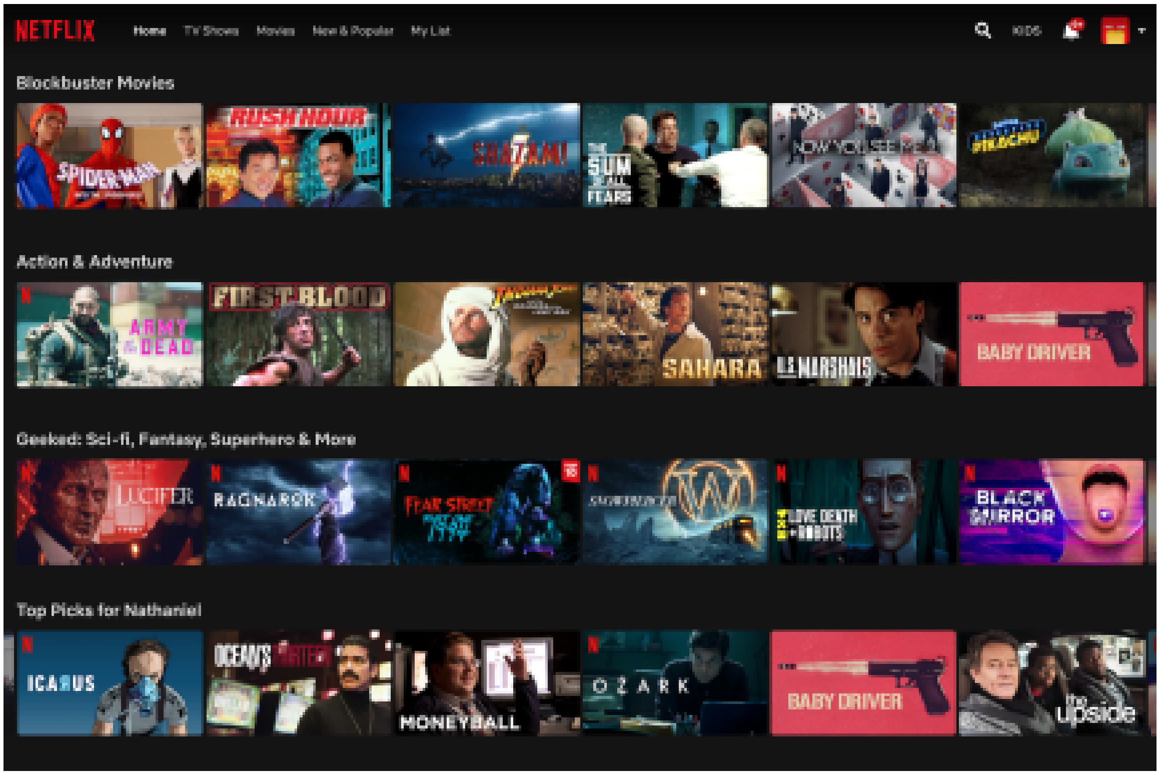
PREAMBLE

Netflix, at one time just an online DVD rental service, has become a titan in the entertainment industry. While predominantly a streaming service, Netflix has also become well-known for its original programming such as the *Stranger Things* television series, the Oscar-nominated film *Marriage Story*, and the ludicrous documentary series *Tiger King*.

The success of Netflix is due, in part, to their well-known data-driven culture. Enmeshed within this culture is a strong appreciation for, and exploitation of, designed experiments. Netflix's home-grown ABlaze experimentation platform is well-known in the industry for its sophistication and the "wins" it has helped them achieve. It is perhaps unsurprising, then, that Netflix is a leader in online-experimentation. Though not recent, this job ad from 2016 for a Senior Data Scientist illustrates the organization's experimental maturity. In this role, you would "design, run, and analyze A/B and multivariate tests", "analyze experimental data with statistical rigor", and "adapt existing methods such as Response Surface Methodology (RSM) to online A/B testing".

In this project you will embark on a Netflix-inspired experimental journey with a hypothetical problem and a web-based response surface simulator.

THE PROBLEM



In this project you will be concerned with optimizing the www.netflix.com homepage by way of minimizing *browsing time*. For those unfamiliar with Netflix, a screenshot of the homepage is included above. As is depicted in the screenshot, the homepage is laid out in a grid system in which movies and TV shows appear as tiles with rows differing with respect to some categorization. Though not depicted in the screenshot, when one hovers their mouse over a tile, its size is enlarged and a preview of the show/movie is automatically played in the enlarged window.

When faced with so many viewing options, Netflix users often experience choice-overload and can be overcome by a psychological phenomenon known as decision paralysis. The problem is that it becomes harder to make a decision, and it takes longer to make a decision, when faced with a large number of options to choose from. Decision paralysis negatively impacts Netflix because a user may become overwhelmed by all of the options and fatigued by the prospect of making a choice, and may ultimately lose interest and not watch anything.

To overcome this, Netflix tries to help you choose what to watch, and by a variety of mechanisms tries to help you choose quickly. Of relevance is browsing time – the length of time a user spends browsing (as opposed to watching) Netflix. Ideally, browsing time and, in particular, average browsing time would be small. In this project you will conduct a series of experiments to learn *what* influences browsing time and *how* that may be exploited in order to minimize average browsing time. There are infinitely many things that likely influence the amount of time someone spends browsing Netflix, but just four factors will be explored in this project. Each is related to the “Top Picks For. . . ” row of the Netflix homepage. This row contains recommendations algorithmically curated for the specific user.

- **Tile Size:** The ratio of a tile’s height to the overall screen height. Note the tile’s aspect ratio is fixed so changing this factor changes the size of the tile, but not its shape. Smaller values correspond to a larger number of tiles visible on the screen, and larger values correspond to fewer visible tiles.

- **Match Score:** A prediction of how much you will enjoy watching the show or movie, based on your viewing history. This is recorded as a percentage, with larger values indicating a higher likelihood of enjoyment.

- **Preview Length:** The duration (in seconds) of a show or movie’s preview.

- **Preview Type:** The type of preview that is autoplayed.

The table below summarizes the region of operability for each of these factors, and the default values they take on when not being experimented with.

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

[a] For purposes of experimentation Match.Score must be an integer

[b] For purposes of experimentation Prev.Length can only be changed in increments of 5 seconds

[c] TT stands for *teaser/trailer* and AC stands for *actual content*

Through a series of experiments you will seek to determine which of these factors significantly influences browsing time, and you will attempt to find an optimal configuration of them that minimizes expected browsing time. You will do this by interacting with a web-based simulator, into which you will submit experimental designs and out of which you will receive response observations.

The remainder of this document provides guidelines for using the simulator, an overview of the sequential experimentation process you will undertake, and a description of the deliverable that you must submit. An outline of the marking scheme is included as an Appendix to make clear my expectations and to make transparent the manner in which you will be graded.

THE EXPERIMENTS

PHASE I: Factor Screening

Use a two-level experiment (i.e., 2^K factorial or 2^{K-p}fractional factorial) to determine *which* factors significantly influence the response. A factor deemed insignificant can be ignored in all subsequent phases of experimentation.

You will experiment with three factors: Tile.Size, Match.Score, Prev.Length. The *low* and *high* levels of these factors (for **this** experiment) are shown below.

Factor	Low	High
Tile.Size	0.1	0.3
Match.Score	80	100
Prev.Length	100	120

Using the data collected from your two-level experiment, determine which factors significantly influence browsing time. Be sure to include formal hypothesis tests and main effect plots in your analysis.

PHASE II: Method of Steepest Descent

Considering only those factors deemed to significantly influence browsing time in PHASE I, perform a *method of steepest descent* analysis to move from the initial region of experimentation toward the vicinity of the optimum. Note that this may require intermediate two-level designs to reorient toward the optimum. You will find tests for curvature and a plot of average browsing time vs. step number useful.

NOTE: the initial region of experimentation is *not* in the vicinity of the optimum, and embarking down the path of steepest descent is necessary. You may use this fact without justification.

PHASE III: Response Optimization

Once you are confident that you are in the vicinity of the optimum, conduct a central composite design and use a second-order response surface model to identify the location of the optimum (i.e., the factor levels that minimize expected browsing time). Report the estimate and a 95% confidence interval for the expected browsing time at this location.