Experimental Design Final Project

Shutong Zhang 20764248

Executive Summary

We believe that Netflix's homepage can be optimized by minimizing average browsing time users spend before watching content. Long browsing time is expected to be linked with shorter watch time and lower customer satisfaction. In this project we conducted a series of experiments following response surface methodology to find the shortest expected browsing time. Three factors identified from the "Top Picks For..." suggestion row are selected to explain browsing time and they are preview length, a piece of content's match score with the user, and the size of a thumbnail.

First, a factor screening experiment is conducted to identify factors that significantly influence the response. A 2^3 factorial design was incorporated and we found that preview length and match score are significant. Next, we used the method of steepest descent to move towards the optimum vicinity. We descended six steps in the response surface following a line that intercepts the origin with slope close to one. The 5th step produced the minimum average browsing time out of the six steps so factors were releveled with step 5 being the center point. A test for curvature followed and showed evidence for curvature in the vicinity. Finally, a central composite design experiment was conducted to approximate a second-order response surface. An optimum was located.

Average browsing time is expected to be minimized at 9.991 minutes when preview length is 60 seconds and match score is 80. A 95% confidence interval of the response is [9.502, 10.481]. From the experiment results we conclude that browsing time can be minimized when preview length decrease from the default 75 seconds to 60 seconds and suggested content's match score decrease from the default 95 to 80. We have learned that shorter previews in the 60 second range reduces browsing time compared to longer previews. Additionally, a match score close to perfect actually increases browsing time. Combining the findings together, content suggestions for users should be a title where the user is somewhat familiar with but would not otherwise think to watch (match score of 80). The preview should be a quick trailer that gives users a taste of the content. Updating the internal suggestion algorithm to follow the findings would most likely shorten browsing time and reduce any negative impact caused by long browsing time in the future.

Introduction

The Problem

Netflix is a video streaming website where users can access over thousands of titles with just one monthly subscription. While it is no doubt that consumers want to have access to as many shows as possible, the large catalog can generate several problems. We will focus on the browsing time consumer behaviour in this project. In our context, browsing time is defined as the time Netflix users spend browsing and searching for content before picking something to watch. When there are many titles to choose from, users might experience a psychological phenomenon known as decision paralysis. In short, a large library makes it harder for users to decide on what to watch. Persistent decision paralysis negatively impact Netflix because when a user struggles to find the next content to watch, they might become frustrated and end up not watching anything. The negative impact will appear in the form of reduced active watch time and perhaps even lower customer satisfaction. We should aim to minimize browsing time.

One of the first place where users go when they want content recommendation is the "Top Picks For..." row. We will focus on using factors related to this row to optimize browsing time. Three factors are initially identified:

- **Preview Length**: The duration (in seconds) of a show's preview. Duration must be multiples of 5 seconds.
- **Match Score**: A prediction of how much you will enjoy a show based on past viewing history. This is recorded as a percentage where a score of 100 means a perfect match with the user.
- **Tile Size**: The ratio of a tile's height to the overall screen height. The aspect ratio is fixed regardless of the tile size.

Factor	CodeName	RegionofOperability	Default
Preview Length	Prev.Length (x1)	[30,120]	75.0
Match Score	Match.Score (x2)	[0,100]	95.0
Tile Size	Tile.Size (x3)	[0.1,0.5]	0.2

The project will explore the relationship between these three factors and the response, coded `Browse.Time`. We will attempt to minimize the metric of interest, average `Browse.Time` by conducting a series of experiments. Data used in all experiments will be collected from a simulator developed by Professor Nathaniel Stevens.

Response Surface Methodology

This project will follow response surface methodology (RSM) to obtain the optimal response. RSM is a method used to explore relationship between multiple explanatory variables and response variables. The philosophy behind RSM is sequential experimentation. We will use what we have learned from one experiment to determine the design of the next one. Instead of conducting one broad, costly experiment like a multi-level factorial experiment with lots of factors, RSM is more efficient in that all of the conditions collected is useful towards locating to the optimum.

Generally speaking, the first phase of RSM is factor screening. In this phase we hope to determine which explanatory variables significantly affect the response. The insignificant variables are then dropped in sequential experiments. Factor screening helps reduces the conditions required in later experiments. Experimental designs for factor screening include the 2^K factorial and 2^{K-p} fractional factorial designs. In phase two of RSM we want to locate the vicnity of the optimum. In this project, the method of steepest descent will be used. We will alternate between traversing a first-order approximation of the response surface and tests

for curvature until we are in the vicinity of the optimum. In the final phase, we will conduct a central composite design in the vicinity of optimum to approximate a second-order response surface. We will locate the optimum on the approximated surface.

RSM can be an efficient method when we want to optimize a response variable. We must note that we can only expect the second-order approximation to be adequate in a small localized region that contains the optimum. This is why phase two must be conducted before phase three. A convention of RSM is to transform continuous covariate values from their natural scale to a coded scale as defined by the formula:

$$x = \frac{U - (U_H + U_L)/2}{(U_H - U_L)/2}$$

where U_H and U_L represent the covariate's high and low levels respectively. In this project, conditions will be denoted in coded scale unless otherwise specified.

Phase I: Factor Screening

The first experiment we will conduct is a factor screening experiment. We have decided to try to explain the response variable, `Browse.Time`, by using three explanatory variables, `Prev.Length`, `Match.Score`, and `Tile.Size`. We want to first know which covariates significantly influence the response. If an insignificant covariate is present, we would consider not controlling for said variable in further experiments to save experimental resources. Factor screening is based on the Pareto Principle where only a few factors will actually significantly influence the response.

We will run a factor screening experiment using a 2^3 factorial design. The response variable is `Browse.Time` and the design factors are `Prev.Length`, `Match.Score`, and `Tile.Size`. Two levels labeled high and low are selected within the region of operability for each of the three design factors. The intentional difference between the high and low levels is to give each factor a fair chance of showing their effects. A summary of each design factor is noted on Table 1. There will be $2^3 = 8$ experimental conditions derived from the unique combinations of the design factors' level. Results of n = 100 units will be simulated per condition for a total of 800 experimental units.

Table 2: 2^3 factorial design

Factor	Low	Center	High
Prev.Length (x1)	100.0	110.0	120.0
Match.Score (x2)	80.0	90.0	100.0
Tile.Size (x3)	0.1	0.2	0.3

We will use a OLS linear regression model to check significance. The model of interest contains all the main effects and interactions for a total of 8 parameters including the intercept. Each of the effects will be estimated by one parameter. Main effects will correspond to $\beta_1, \beta_2, \beta_3$, two-way interactions will correspond to $\beta_{12}, \beta_{13}, \beta_{23}$, and three-way interaction will correspond to β_{123}. We can determine significance from a series of t-tests. For each effect, $H_0 : \beta_i = 0$ tests whether the effect is significant or not.

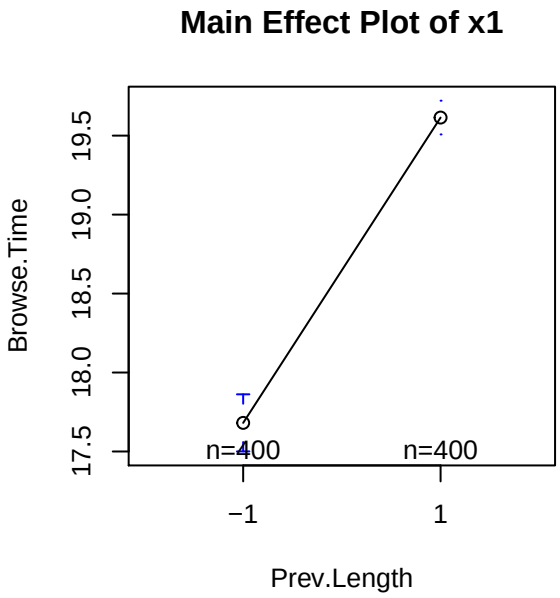
```
##   beta_0   beta_1   beta_2   beta_3  beta_12  beta_13  beta_23 beta_123
##   0.0000   0.0000   0.0000   0.4195   0.0000   0.4608   0.6225   0.2481
```

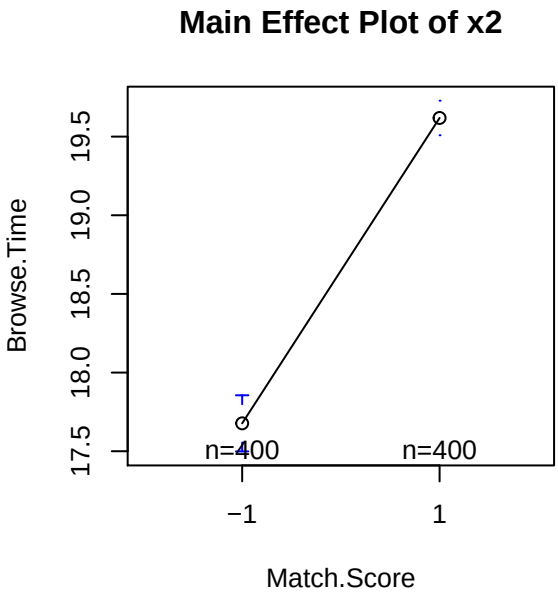
The output above are p-values of the said hypothesis tests rounded to 4 decimal places. At a $\alpha = 0.05$ significance level, we reject null hypotheses that β_1, β_2, and β_{12} are 0. Hence we conclude that `Prev.Length` and `Match.Score` are the only significant factors. We will not experiment with `Tile.Size` in upcoming experiments and as a result, saving experimental conditions in later experiments.

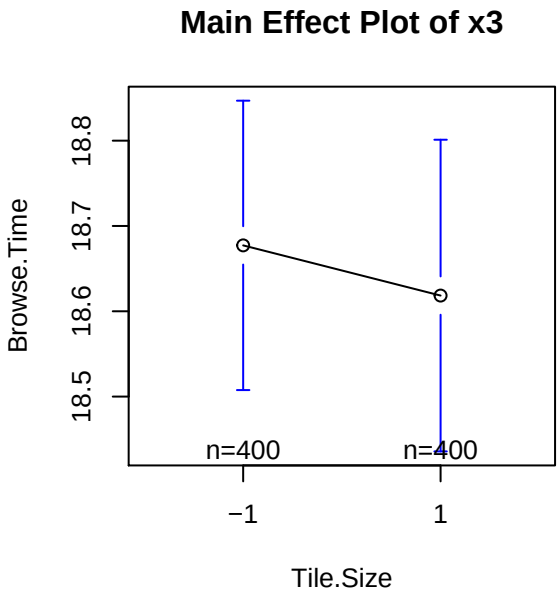
From the plots below, we confirm what we have learned from the hypothesis tests. The slopes on the `x1`, `x2` main effect plots appears to be significantly different from 0. On the other hand, `x3` main effect plot might have slope 0 if we account for the variance of the response at each level. From the `x1:x2` interaction plot we see that the interaction effect is also significant. `Browse.Time` is minimized when `Prev.Length` and `Match.Score` are both at their low levels.

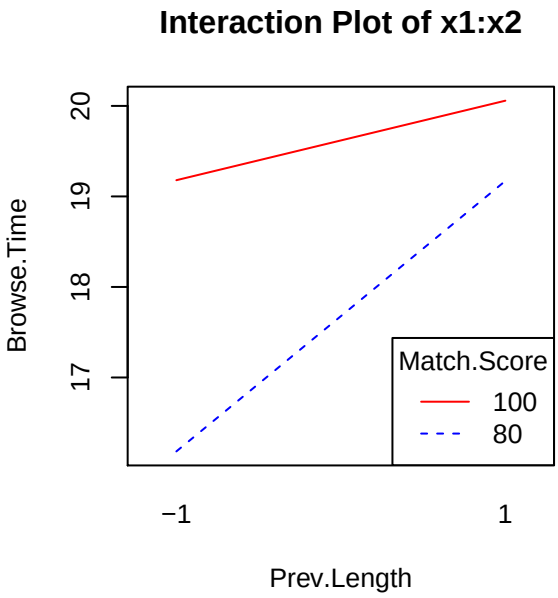
```
# main effects
```

```
## Prev.Length Match.Score   Tile.Size
##   1.9331781   1.9410631  -0.0588468
```









Alternatively, we could have used a 2^{3-1} fractional factorial design. If we aliased `x3` with a two-way interaction, we could have conducted a similar factor screening experiment with just $2^{3-1} = 4$ experimental conditions. However, the cost of aliasing is confounding. Every main effect will now be jointly estimated with a two-way interaction effect. Considering we only have 3 design factors to begin with, 2^3 - $2^{3-1} = 4$ extra experimental conditions will allow us to avoid any ambiguity. Especially if we consider the principle of effect sparsity which states that main effects and two-way interactions are the most likely to influence the response, we forewent any risk of misinterpretation and chose a 2^3 factorial design to seperately identify each effect.

Phase II: Method of Steepest Descent

The next problem of interest is locating the region in the response surface that contains the minimum response. The response surface that we will explore is the relationship between `Prev.Length`, `Match.Score`, and expected `Browse.Time`. We were given the information that the region experimented in phase 1 does not contain the optimum. Therefore the first thing to investigate is the direction towards the optimum. We will estimate a first-order response surface using a linear regression model containing only the main effects of `Prev.Length` and `Match.Score`. Then using the parameter estimates, we can define the gradient of the first-order response surface as $g = [\hat{\beta}_1 \quad \hat{\beta}_2]$. This will be the direction of steepest decrease on the fitted surface and we will traverse in steps following the gradient starting from step 0, $x_0 = (0,0)$.

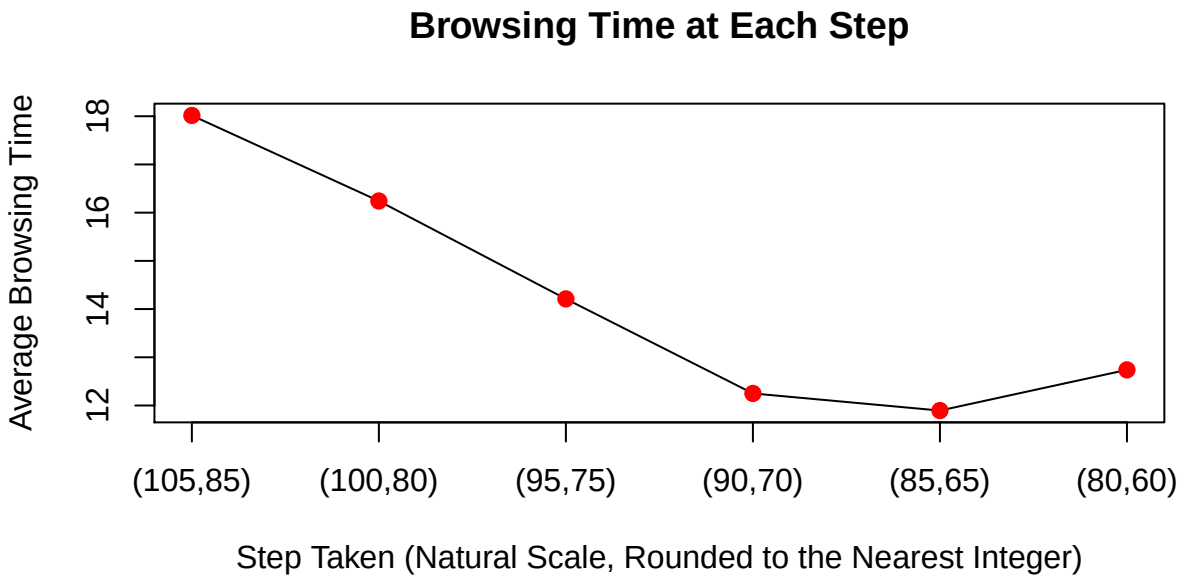
To determine where step n is, we will use the formula:

$$x_n = x_{n-1} - \lambda g, \quad where \;\; \lambda = \frac{\Delta x_1}{|\hat{\beta}_1|}$$

λ is the fixed step size we will take from x_{n-1} to x_n. It is defined based on `Prev.Length` because it is the design factor hardest to manipulate. We will consider $\Delta x_1 = 0.5$ since it conveniently represents a 5 second change in `Prev.Length`. From here we will repeat an algorithm that efficiently uses experimental resources until we find the first minimum response on the first-order response surface. The algorithm is as follows:

1. Calculate x_n and collect simulated data (n = 100). Compute average `Browse.Time`.
2. Repeat step 1 using data from condition x_{n+1} until average `Browse.Time` stops decreasing.

The nth condition that produced the smallest average `Browse.Time` will become the new center point. Low and high levels for `Prev.Length` and `Match.Score` will be releveled so that x_n becomes the center point. With the center point and 4 combinations of unique factor levels, we will test for curvature at (0,0). A reasonably wide range between the low and high levels should be considered to account for where curvature can exist. However, if the range is excessively wide, a second-order response surface approximation might not be accurate according to Taylor Series. The test for curvature uses a linear regression model to check if there is a pure quadratic effect. If there is evidence of curvature, we are in the vicinity of the optimum.



From the estimated first-order surface we get $\lambda = 0.5172829$ and $g = (0.966589, 0.9705316)$. Their product is $\lambda g = (0.5, 0.502)$. Starting from our original center point at (0,0) we took 6 steps in total to find the condition that resulted in minimum average `Browse.Time`. Step 5, the (-2.5, -2.51) condition ((85, 64.9) in natural scale) returned the shortest average `Browse.Time` at 11.90 minutes. Our re-centered coded scale in this new region is defined in Table 2:

Table 3: Re-centered Levels

Factor	Low	Center	High
Prev.Length (x1)	70	85	100
Match.Score (x2)	55	65	75

Four two-level factorial conditions were simulated (n = 100) for a total of 400 experimental units. Using the newly collected data and the center point condition collected from step 5, we built a linear regression model that tested for curvature. A center point indicator variable denoted x_{PQ}, x_1, x_2 and their interaction x_1x_2 were included in the regression model. We formally tested $H_0 : \beta_{PQ} = 0$ using a t-test to see if the pure quadratic effect is significant.

```
# p-values of significance of coefficients in the model
```

```
##             (Intercept)             Prev.Length             Match.Score
##            0.000000e+00            2.542383e-88            3.888043e-65
##                      pq Prev.Length:Match.Score
##            3.965489e-65            2.424104e-90
```

We rejected H_0 at a $\alpha = 0.05$ significance level. There is strong evidence for curvature somewhere in the response surface. We can say we are in the vicinity of the optimum response. In the next phase, we will conduct a central composite design to identify the location of the optimum.

Phase III: Response Optimization

In the final stage, we want to locate the optimum. We will need to find the value of `Prev.Length` and `Match.Score` that minimize the expected `Browse.Time` in the region defined in Phase II. The low and high levels of the factors were chosen to cover a reasonable amount of the response surface. The region was designed to not be excessively small nor large which will lead to unhelpful surface approximations. The condition that minimizes the response is referred to as the stationary point. We will choose a central composite design in order to fit a full second-order response surface model. There are three types of experimental conditions in a central composite design:

1. **Two-level factorial Conditions**: There are four in total and they are already collected in Phase II. They are $(\pm 1, \pm 1)$
2. **A center point Condition**: This was also collected in Phase II. It is $(0, 0)$
3. **Axial Conditions**: There are four in total and they are $(\pm a, 0)$, and $(0, \pm a)$.
    - `a` is an arbitrary value determined by the experimenter. In this experiment, a = 1 is chosen becauase of practical concerns. Setting a = 1 allows us to only consider 3 levels for every factor. Additionally, it satisfies the mutiples of 5 seconds constraint on `Prev.Length`.

In total 9 experimental conditions each with n=100 randomly simulated units will be used for the central composite design. These conditions will allow us to fit a linear regression model with main effects, two-way interactions, and quadratic effects.

```
##
## Call:
## lm(formula = Browse.Time ~ Prev.Length + Match.Score + Prev.Length *
##     Match.Score + I(Prev.Length^2) + I(Match.Score^2), data = ccd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6036 -0.6799 -0.0071  0.7143  2.8697
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             11.97887    0.07609  157.44   <2e-16 ***
## Prev.Length              1.37440    0.04167   32.98   <2e-16 ***
## Match.Score             -1.10059    0.04167  -26.41   <2e-16 ***
## I(Prev.Length^2)         0.94332    0.07218   13.07   <2e-16 ***
## I(Match.Score^2)         1.08427    0.07218   15.02   <2e-16 ***
## Prev.Length:Match.Score  1.24230    0.05104   24.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.021 on 894 degrees of freedom
## Multiple R-squared:  0.7563, Adjusted R-squared:  0.7549
## F-statistic: 554.8 on 5 and 894 DF,  p-value: < 2.2e-16
```

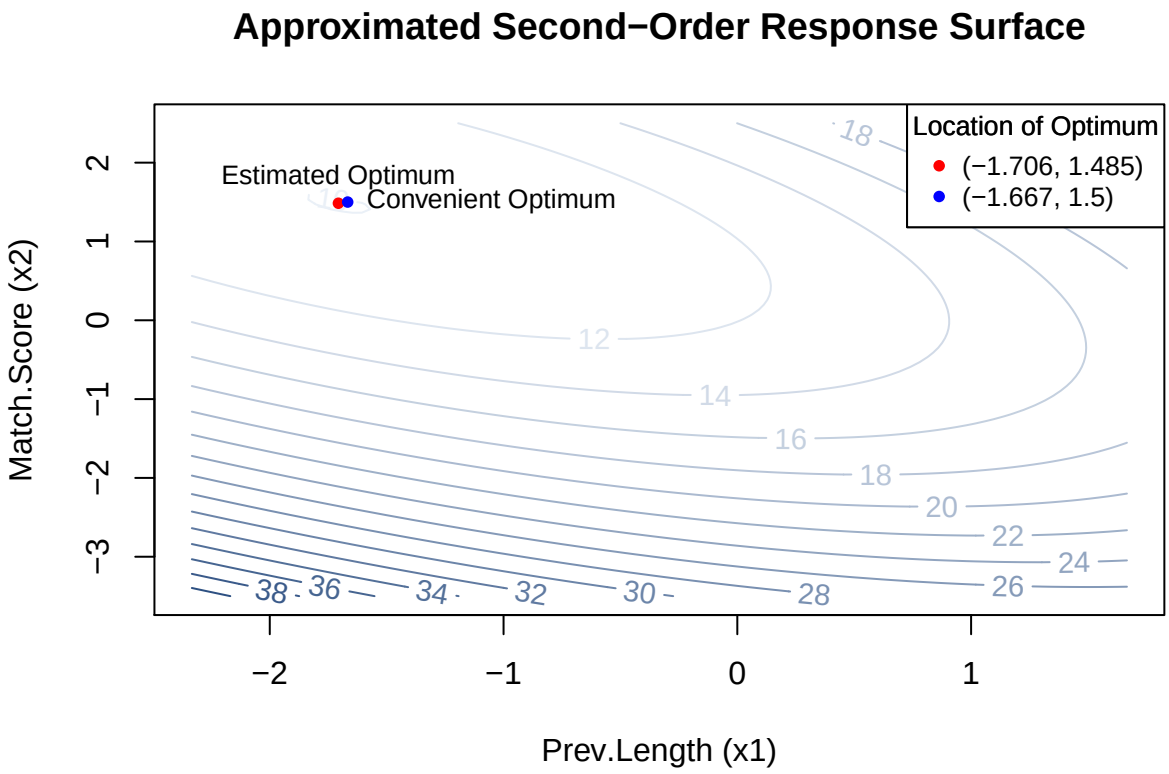
We fitted the full second-order model and obtained the summary above. The model appears to explain the response well with respect to $R^2 = 0.7563$. Also, the t-tests suggests that all of the coefficients are significantly non-zero at a $\alpha = 0.05$ significance level. The model is a valid fit based on the residual summary. The quantiles closely resembles the standard normal distribution. We are confident that `model.ccd` is a good second-order estimate of the response surface. To obtain the minimum expected response, we will solve for the stationary point using the formula:

$$\mathbf{x}_s = -\frac{1}{2}\mathbf{B}^{-1}\mathbf{b} \qquad \textit{where} \quad \mathbf{b} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \quad \textit{and} \quad \mathbf{B} = \begin{bmatrix} \hat{\beta}_{11} & \frac{1}{2}\hat{\beta}_{12} \\ \frac{1}{2}\hat{\beta}_{12} & \hat{\beta}_{22} \end{bmatrix}$$

The optimal expected response is given by:

$$E[\hat{Y}] = \hat{\beta}_0 + \frac{1}{2}\mathbf{x}_s^T\mathbf{b}$$

We get $\mathbf{x}_s$ = (-1.706, 1.485) which translates to (59.404, 79.851) in the natural scale. The expected `Browse.Time` at the stationary point is 9.989 minutes. The 95% Prediction interval of the expected response at the stationary point is [9.488, 10.490]. Since `Prev.Length` has to be multiples of 5 seconds, we will consider the expected response at $\mathbf{x}$ = (60, 80) in the natural scale instead. The expected `Browse.Time` at the convenient $\mathbf{x}$ is 9.991 minutes. The 95% Prediction interval of the expected response at the new $\mathbf{x}$ is [9.502, 10.481]. Since the results are comparable, we will conclude that the expected `Browse.Time` is minimized when $\mathbf{x}$ = (60, 80). The contour plot below plots the approximated response surface and the location of the optimum.



To conclude the project, browsing time can be minimized when preview length decrease from the default 75 seconds to 60 seconds and suggested content's match score decrease from the default 95 to 80. We have learned that shorter previews in the 60 second range reduces browsing time compared to longer previews. Additionally, a match score close to perfect actually increases browsing time. Combining the findings together, content suggestions for users should be a title where the user is somewhat familiar with but would not otherwise think to watch (match score of 80). The preview should be a quick trailer that gives users a taste of the content. Updating the internal suggestion algorithm to follow the findings would most likely shorten browsing time and reduce any negative impact caused by long browsing time in the future.