**Summary**

We used regression analysis to determine what factors have the greatest predictive power for movie ratings and then used Solver to optimize the equation we were given by Minitab.  This new equation fairly accurately predicts a user's rating for a particular movie, in this case Independence Day.  Given more time, we could try different movies, genres, dependent, independent variable etc to fine tune and improve our process.

**Exposition**

The intent of this project was to formulate an equation which would allow us to accurately predict the ratings, 1-5, that a user would give a particular movie. The idea for the project was derived from the Netflix Challenge.

The Netflix Challenge is an ongoing competition held by NETFLIX to find the best algorithm that predicts user ratings. The goal of the competition is to beat Netflix’s current algorithm, Cinematch, by improving the root mean squared error (RMSE) of a target list of ratings by 10%. The grand prize of the competition is $1,000,000. Ratings include 17,700 films rated by 480,000 unique users, however not all users have rated all movies, creating “holes” in the data.

We felt this project was interesting and challenging because of its real world application of Decision Models concepts, its broad appeal and multiple applications, the extensive data and direction provided by Netflix and of course because of the $1,000,000 prize!

Netflix provides ratings data for 17,700 movies rated by 480,000 Netflix customers. In addition we collected attributes data such genre, box office, release date, etc of 30 target movies. Then we selected a subset of users who rated the 30 target movies.

Using this information we ran several regressions with the independent variables being movie attributes and the dependent variables the ratings of several “representative” movies from a group of users. However these regressions did not provide relevant information on which attributes were most predictive.

Next we ran several regressions which included movie attributes as well as the ratings of 9 representatives movies. These 9 movies represented nine genres -

* Action/Adventure
* Comedy
* Drama
* Epic/Historic
* Family
* Science Fiction
* Musical
* Horror
* Western

Based on these regressions, prior ratings appear to be more meaningful than movie attributes. Also we determined that including 1 movie per genre yields an R-Sq of 51.6%, while 2 movies per genre yields an R-Sq of 90.8%. To test the regression, we used the ratings of 30 new, randomly selected users. The one negative with this approach is that the regression is movie specific.

Next we used Solver to optimize our final regression. Since Netflix measures Challenge submissions by Root Mean Squared Error (RMSE), we decided to make the objective of the solver model to minimize RMSE of our sample ratings, while the model’s variables were the regression factors for the 18 movies ratings (2 per genre) and the only constraints was that computed ratings must be ≤ 5. This new equation fairly accurately predicts a user's rating for a particular movie, in this case Independence Day.

The biggest challenges in the project included:

* Format of the data provided by Netflix. We were given 17,700 text files containing ratings data.
* Not all 17,700 titles have a statistically significant number of ratings.
* Some users only rated 1 movie while one user rated all 17,700 movies.
* Each movie requires a unique model.

However, given more time and a more complete database package, we feel we could make significant progress.