CineMatch

Andy Samant, Arjun Shah, Vladimir Sekiguchi, Victor Chu

12/14/2016

==========

PROBLEM STATEMENT

We are attempting to predict whether users like or dislike certain movies based on movie review data, movie information data, and user demographic data. Movie platforms and distributors such as Netflix or Amazon seek to provide their customers with a movie recommendation service-a method of assisting a user in finding new material to watch or buy based on that user's personal tastes, as determined by previous indulgences and feedback (reviews). In doing so, such services hold customer interest. Consequently, these companies, whose business models rely on subscription and viewership, may maintain brand loyalty and dependency from their consumer base and continue to thrive.

==========

DATA DESCRIPTION

The data files we used contained 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users who joined MovieLens in 2000. There were three separate files: a ratings file, a users file, and a movies file. The ratings file ("ratings.dat") contained UserIDs ranging between 1 and 6040, MovieIDs ranging between 1 and 3952, ratings made on a 5-star scale (whole-star ratings only), and timestamps. Each UserID in the dataset was ensured to map to at least 20 ratings. The users file ("users.dat") contained UserIDs ranging between 1 and 6040, gender denoted by a "M" for male and "F" for female, age chosen from specified yearly ranges (1: "Under 18", 18: "18-24", 25: "25-34", 35: "35-44", 45: "45-49", 50: "50-55", 56: "56+"), and occupation chosen from specified choices (0: "other" or not specified, 1: "academic/educator", 2: "artist", 3: "clerical/admin", 4: "college/grad student", 5: "customer service", 6: "doctor/health care", 7: "executive/managerial", 8: "farmer", 9: "homemaker", 10: "K-12 student", 11: "lawyer", 12: "programmer", 13: "retired", 14: "sales/marketing", 15: "scientist", 16: "self-employed", 17: "technician/engineer", 18: "tradesman/craftsman", 19: "unemployed", 20: "writer"). The movies file ("movies.dat") contained MovieIDs ranging between 1 and approximately 3,900, title as provided by the IMDB (including the year of release), and pipe-separated genres chosen from a specified list of genres (Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western).

==========

DATA PREPROCESSING

Once the three data files were imported, their columns were labeled and they were merged. New columns were added to the data frame: a column was added with quasi-boolean values (0's and 1's) for each genre to split up the pipe-separated genre column from the ratings.dat file, and columns were added to give aggregate mean values for ratings based on MovieID, ratings based on UserID, and ratings based on UserID per genre. UserIds for which aggregate mean values per genre were unable to be retrieved (due to that user not having rated any movies of that genre) were assigned the value 0 as a default. Columns that contained information that seemed irrelevant or inhibitory to our cause (such as the timestamp, zip-code, and title columns) were nullified to save time and processing power. All the columns containing aggregate mean ratings were factored according to their numerical rating (factors between 1 to 5), and the rating column was factored as a "like" or "dislike" depending on its values (ratings greater than 0 and less than or equal to 3 were factored as "dislikes" and ratings greater than 3 and less than or equal to 5 were factored as "likes") as we sought to simplify our results into "likes" or "dislikes" for the sake of accuracy; we expect a higher hit rate while predicting a binary value rather than forecasting a numerical rating from 1 to 5. We decided that the data did not need balancing, as there was a relatively even spread of rating between the two factors-approximately 425000 dislikes to 575000 likes. It should be noted that we frequently ensured that our columns were purely factored and no missing values existed before proceeding. Finally, we divided our ratings data with 70% of the data going to the training dataset and the remaining 30% going to the testing dataset.

==========

MACHINE LEARNING APPROACH

We applied this techniques to our training data sets, 70% of our datasets, to produce classification models that would give us an answer to whether a user likes or dislikes a given movie. After we developed these models, we applied it to the remaining 30% of our testing data sets to get our performance results (accuracy, sensitivity, specificity, negative predictive power, and positive predictive power; all of which will be elaborated on in the results.

Generalized Linear Model:

This is a logistic regression technique that we used to measure the relationship between our categorical dependent variable (our outcome that determines whether an arbitrary user will like or dislike a movie) and our 43 independent variables (our predictor variables) by using a cumulative logistic distribution. We chose this method to use as a control to show how well the other two algorithms perform to this one, which we knew from the start would present a not-so-great predictor model.

Random Forests:

```
    This machine learning algorithm starts off by growing a large number of c
lassification/decision trees by sampling N cases with replacement from our tr
aining set. Classification/decision trees are essentially predictive models t
hemselves that map observations about users and movies from our dataset to co
nclusions about the user's binary opinion on the movie. After that, we take s
amples of size sqrt(M), where M is the total number of predictor variables at
each node, in order to determine how many variables are candidates to be used
for splitting that node. Then we grow each tree fully without pruning, or cut
ting out nodes from the bottom. The nodes at the bottom are assigned a class
based on the case that shows up the most in that node. For all new classifica
tion cases, we can send them down the tree to the bottom nodes and take a maj
ority vote to determine the outcome variable (like or dislike). We chose this
algorithm because we knew it would perform well since it is meant for classif
ying for binary decisions, and it will generate highly correlated trees which
will not reduce the variance of our results by too much.

##
## Call:
##  randomForest(formula = vars, data = train, dna.action = na.roughfix,
ntree = 10, importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 10
## No. of variables tried at each split: 6
##
##         OOB estimate of  error rate: 31.03%
## Confusion matrix:
##         Dislike   Like class.error
## Dislike  171382 123254   0.4183263
## Like      91861 306678   0.2304944
```

Boosting:

```
    This method begins by assigning equal weights to each case out of N cases
, which is 1/N. We then fit a classifier, which is weak because the accuracy
is just above 50% and just better than taking a 50/50 guess for whether a use
r likes a movie or not. For the misclassified cases, we give them a greater w
eight and refit a new classifier. We repeat this process around (2-5) * M tim
es, where M is the number of our predictor variables. We then combine the M c
lassifiers by averaging them out and giving a greater weight to the classifie
rs who perform better and had better accuracy. We chose this algorithm becaus
e we knew it would build a strong classifier from a set of weak classifiers w
hich would create a good model for our movie predictions.
```

==========

RESULTS

Let's keep in mind, the purpose of these models is to predict whether or not a given user will like a given movie. We need to look at three types of predictors: information about the user, information about the movie, and information about the interaction between the user and the movie.

Only some of the demographic predictors mattered. The occupation-based factors are particularly interesting. These were some of our most statistically significantly factors. Artists, craftspeople, customer service workers, programmers, scientists, and unemployed people were significantly more dissatisfied with their movies than the other occupations. The effects were strongest for artists and the unemployed. Farmers and K-12 students were more likely to give higher ratings to their movies. In terms of ages, older users gave lower ratings. The effects are small or nonexistent for users aged 34 or under. However, the effect on the rating is negative for older groups, with the standard error falling further and further below zero as the age level increases. However, after age 56, age has a negligible effect on ratings. Men are also expected to have slightly higher movie ratings than women. We also have data on the users' existing movie ratings. Users were also far more likely to give a movie a positive review if their average rating, across all movies, was high. The higher the mean rating, the more powerful the effect, for obvious reasons. This could mean that certain groups are naturally more optimistic about movies than others and tend to give them higher ratings. They might enjoy almost anything that they watch. Other viewers are tougher because they tend to dislike many movies. While these factors don't directly help us recommend movies to users, it does help us put their previous reviews in context so that the model can better examine the other relevant factors. Adjusting for demographic factors is important.

Next, we need to look at factors pertinent to the movie itself. In this model, the mean rating for the movie had a negligible effect. One of the most important predictors was the genre of the movie. All of these coefficients were positive and very significant. Genres that received higher ratings include: animation, drama, comedy, crime, fantasy, musicals, mysteries, westerns. Genres that tended to receive lower ratings included: documentaries, horror, and war. The probability that these coefficients were less than the absolute value of the z-statistic (the probability of significance) was 2e-16, or as significant as they can possibly to be. These are incredibly important predictors of any given rating. From this information, we know which genres are the most popular among users.

However, we lastly have the interaction variables. A user's typical rating of a horror movie matters most when the movie we are trying to match them with is a horror movie. So we tested the interaction of mean genre ratings and the Boolean of whether or not it is that genre. Every single one of these predictors was also as significant as possible and had a very negative effect on movie ratings, except for single star ratings, which had a negligible effect. The single star ratings are likely negligible, probably because there were so

few of them. The effects were strongest when the mean ratings were a two or three. The coefficients were noticeably closer to zero for mean ratings of four stars, but still negative. Ratings of five stars were not included in the results because these variables are factors, and one level of the factors must always be left out. We can figure out their effect though. Logically, if the mean user rating was between one and four (so none of their negative coefficients would be included) the ratings would likely be more positive. Therefore, the effect of a five star genre mean rating is positive. This makes sense. Users that give high ratings to action movies are likely to enjoy other action movies. Interestingly, giving a genre an average rating of four stars or less (even an average of 3.9 stars) means that you are unlikely to enjoy any movie of that genre. Again, these were some of the most important predictors in the GLM model.

The GLM model was an imperfect predictor model. The confusion matrix shows that the model too often predicted that users would dislike the movie. There were many false negatives but very few false positives. This gave the model near perfect specificity (99% of the time that users disliked a movie, the model would say so) and positive predictive power (96% of the time that the model predicted a user would like a movie, it was accurate). However, it had a very low sensitivity (if users liked a movie, the model would say so only 4.6% of the time) and low negative predictive power (if the model predicted that a user disliked a movie, it would be accurate only 30% of the time). The model was very reluctant to ever predict that a user would like a movie.

```
# GLM results
confusionMatrix(glmpred, test$Rating, positive="Like")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Dislike   Like
##    Dislike   59847 140249
##    Like        283   6826
##
##                Accuracy : 0.3218
##                  95% CI : (0.3198, 0.3238)
##     No Information Rate : 0.7098
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0247
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.04641
##             Specificity : 0.99529
##          Pos Pred Value : 0.96019
##          Neg Pred Value : 0.29909
##              Prevalence : 0.70980
##          Detection Rate : 0.03294
##    Detection Prevalence : 0.03431
##       Balanced Accuracy : 0.52085
```

```
##
##        'Positive' Class : Like
##
```

```
summary(fit.lm)
```

```
##
## Call:
## glm(formula = vars, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0797  -1.0000   0.4589   0.9598   4.0815
##
## Coefficients: (18 not defined because of singularities)
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        -1.345e+01  8.894e+01  -0.151 0.879752
## GenderM                             1.859e-02  7.067e-03   2.630 0.008538
## Age18-24                            6.051e-02  2.824e-02   2.143 0.032152
## Age25-34                           -6.691e-04  2.807e-02  -0.024 0.980987
## Age35-44                           -7.542e-02  2.847e-02  -2.649 0.008083
## Age45-49                           -1.022e-01  2.954e-02  -3.458 0.000543
## Age50-55                           -1.333e-01  2.983e-02  -4.469 7.84e-06
## Age56+                             -5.799e-02  3.204e-02  -1.810 0.070296
## Occupationacademic/educator        -2.583e-02  1.241e-02  -2.081 0.037399
## Occupationartist                   -7.587e-02  1.465e-02  -5.181 2.21e-07
## Occupationclerical/admin            2.468e-02  1.746e-02   1.413 0.157553
## Occupationcollege/grad student     -2.945e-02  1.188e-02  -2.478 0.013202
## Occupationcustomer service         -5.975e-02  2.021e-02  -2.957 0.003108
## Occupationdoctor/health care        3.661e-02  1.665e-02   2.198 0.027952
## Occupationexecutive/managerial      1.409e-02  1.169e-02   1.205 0.228053
## Occupationfarmer                    1.926e-01  5.185e-02   3.714 0.000204
## Occupationhomemaker                -2.691e-02  2.724e-02  -0.988 0.323220
## OccupationK-12 student              1.684e-01  3.030e-02   5.560 2.70e-08
## Occupationlawyer                   -3.480e-02  2.151e-02  -1.618 0.105711
## Occupationprogrammer               -4.278e-02  1.407e-02  -3.040 0.002368
## Occupationretired                   3.892e-04  2.769e-02   0.014 0.988786
## Occupationsales/marketing           2.695e-02  1.490e-02   1.809 0.070461
## Occupationscientist                -6.404e-02  2.019e-02  -3.172 0.001512
## Occupationself-employed            -2.375e-02  1.547e-02  -1.535 0.124823
## Occupationtechnician/engineer       2.279e-03  1.314e-02   0.173 0.862291
## Occupationtradesman/craftsman      -7.386e-02  2.618e-02  -2.822 0.004780
## Occupationunemployed               -8.158e-02  2.481e-02  -3.289 0.001006
## Occupationwriter                   -1.752e-02  1.372e-02  -1.277 0.201693
## Action1                             4.522e-01  2.422e-02  18.673  < 2e-16
## MeanActionRating1                  -1.080e-01  1.394e-01  -0.775 0.438451
## MeanActionRating2                  -1.333e-01  8.454e-02  -1.576 0.114967
## MeanActionRating3                  -5.337e-02  7.370e-02  -0.724 0.468975
## MeanActionRating4                   1.217e-02  7.322e-02   0.166 0.868004
## MeanActionRating5                   1.587e-02  7.390e-02   0.215 0.829989
```

```
## Adventure1                  4.742e-01  3.405e-02  13.924  < 2e-16
## MeanAdventureRating1       -1.047e-01  8.638e-02  -1.212 0.225348
## MeanAdventureRating2       -5.733e-02  4.547e-02  -1.261 0.207362
## MeanAdventureRating3       -6.428e-02  3.340e-02  -1.924 0.054294
## MeanAdventureRating4        1.146e-02  3.283e-02   0.349 0.727110
## MeanAdventureRating5        6.398e-02  3.444e-02   1.858 0.063174
## Animation1                  8.670e-01  4.052e-02  21.397  < 2e-16
## MeanAnimationRating1        2.618e-03  4.045e-02   0.065 0.948399
## MeanAnimationRating2       -3.057e-02  2.760e-02  -1.108 0.267906
## MeanAnimationRating3        5.018e-03  1.558e-02   0.322 0.747423
## MeanAnimationRating4        5.139e-02  1.426e-02   3.605 0.000313
## MeanAnimationRating5        1.092e-01  1.521e-02   7.180 6.98e-13
## Children.s1                 5.692e-01  3.874e-02  14.692  < 2e-16
## MeanChildren.sRating1       6.936e-02  4.707e-02   1.473 0.140633
## MeanChildren.sRating2       1.342e-01  2.627e-02   5.109 3.24e-07
## MeanChildren.sRating3      -1.672e-02  2.010e-02  -0.832 0.405354
## MeanChildren.sRating4      -3.285e-02  1.936e-02  -1.697 0.089741
## MeanChildren.sRating5      -3.371e-02  2.086e-02  -1.616 0.106088
## Comedy1                     5.444e-01  1.854e-02  29.359  < 2e-16
## MeanComedyRating1          -3.160e-01  2.953e-01  -1.070 0.284639
## MeanComedyRating2          -1.761e-01  1.839e-01  -0.958 0.338149
## MeanComedyRating3          -9.572e-02  1.637e-01  -0.585 0.558666
## MeanComedyRating4          -1.021e-02  1.631e-01  -0.063 0.950068
## MeanComedyRating5          -3.581e-02  1.635e-01  -0.219 0.826692
## Crime1                      5.916e-01  2.791e-02  21.194  < 2e-16
## MeanCrimeRating1            1.638e-01  8.146e-02   2.011 0.044295
## MeanCrimeRating2           -5.596e-02  4.532e-02  -1.235 0.216878
## MeanCrimeRating3           -2.517e-02  2.628e-02  -0.958 0.338138
## MeanCrimeRating4           -4.704e-03  2.498e-02  -0.188 0.850632
## MeanCrimeRating5            4.325e-02  2.570e-02   1.683 0.092396
## Documentary1                1.567e+00  8.363e-02  18.742  < 2e-16
## MeanDocumentaryRating1      4.807e-02  2.492e-02   1.929 0.053704
## MeanDocumentaryRating2     -4.721e-02  1.852e-02  -2.549 0.010800
## MeanDocumentaryRating3     -1.363e-01  1.081e-02 -12.612  < 2e-16
## MeanDocumentaryRating4     -6.822e-02  7.702e-03  -8.857  < 2e-16
## MeanDocumentaryRating5      5.961e-03  8.249e-03   0.723 0.469883
## Drama1                      5.906e-01  1.352e-02  43.687  < 2e-16
## MeanDramaRating1            3.187e+00  8.358e-01   3.813 0.000137
## MeanDramaRating2           -8.047e-02  3.263e-01  -0.247 0.805202
## MeanDramaRating3           -3.694e-01  3.115e-01  -1.186 0.235775
## MeanDramaRating4           -3.508e-01  3.111e-01  -1.128 0.259447
## MeanDramaRating5           -2.938e-01  3.111e-01  -0.944 0.345060
## Fantasy1                    1.073e+00  6.619e-02  16.206  < 2e-16
## MeanFantasyRating1          1.641e-02  3.818e-02   0.430 0.667377
## MeanFantasyRating2         -7.136e-02  2.151e-02  -3.318 0.000908
## MeanFantasyRating3         -4.203e-02  1.321e-02  -3.181 0.001466
## MeanFantasyRating4          7.569e-02  1.234e-02   6.134 8.57e-10
## MeanFantasyRating5          1.046e-01  1.543e-02   6.778 1.22e-11
## Film.Noir1                  1.021e+00  4.520e-02  22.587  < 2e-16
## MeanFilm.NoirRating1       -1.517e-01  5.587e-02  -2.716 0.006610
```

```
## MeanFilm.NoirRating2              -1.588e-01  2.803e-02  -5.665 1.47e-08
## MeanFilm.NoirRating3              -1.222e-01  1.321e-02  -9.247  < 2e-16
## MeanFilm.NoirRating4              -1.293e-01  1.059e-02 -12.209  < 2e-16
## MeanFilm.NoirRating5              -5.948e-02  1.101e-02  -5.401 6.63e-08
## Horror1                            1.068e+00  4.347e-02  24.571  < 2e-16
## MeanHorrorRating1                 -4.133e-02  3.740e-02  -1.105 0.269120
## MeanHorrorRating2                  3.685e-02  2.325e-02   1.585 0.112932
## MeanHorrorRating3                 -6.453e-02  1.848e-02  -3.491 0.000480
## MeanHorrorRating4                  4.333e-02  1.811e-02   2.392 0.016763
## MeanHorrorRating5                  4.003e-02  2.059e-02   1.945 0.051818
## Musical1                           8.744e-01  3.715e-02  23.537  < 2e-16
## MeanMusicalRating1                 1.445e-01  4.445e-02   3.251 0.001151
## MeanMusicalRating2                 8.211e-02  2.283e-02   3.597 0.000322
## MeanMusicalRating3                -3.507e-02  1.499e-02  -2.340 0.019269
## MeanMusicalRating4                -2.148e-02  1.388e-02  -1.547 0.121793
## MeanMusicalRating5                 2.681e-02  1.505e-02   1.782 0.074723
## Mystery1                           7.797e-01  4.382e-02  17.793  < 2e-16
## MeanMysteryRating1                -2.094e-02  4.980e-02  -0.421 0.674103
## MeanMysteryRating2                 4.322e-02  2.675e-02   1.616 0.106120
## MeanMysteryRating3                -3.330e-02  1.602e-02  -2.079 0.037654
## MeanMysteryRating4                 5.048e-02  1.477e-02   3.418 0.000630
## MeanMysteryRating5                 4.833e-02  1.589e-02   3.042 0.002350
## Romance1                           5.462e-01  2.396e-02  22.794  < 2e-16
## MeanRomanceRating1                 3.245e-01  1.255e-01   2.585 0.009725
## MeanRomanceRating2                -1.060e-01  6.899e-02  -1.536 0.124584
## MeanRomanceRating3                -2.723e-02  4.870e-02  -0.559 0.576034
## MeanRomanceRating4                 2.301e-02  4.776e-02   0.482 0.629897
## MeanRomanceRating5                 6.075e-02  4.846e-02   1.254 0.209982
## Sci.Fi1                            5.331e-01  3.271e-02  16.297  < 2e-16
## MeanSci.FiRating1                  3.827e-02  7.733e-02   0.495 0.620674
## MeanSci.FiRating2                 -2.655e-02  4.274e-02  -0.621 0.534476
## MeanSci.FiRating3                 -5.291e-02  3.241e-02  -1.633 0.102571
## MeanSci.FiRating4                  2.713e-02  3.203e-02   0.847 0.396849
## MeanSci.FiRating5                  6.242e-02  3.387e-02   1.843 0.065355
## Thriller1                          5.551e-01  2.244e-02  24.738  < 2e-16
## MeanThrillerRating1                4.335e-01  1.637e-01   2.648 0.008088
## MeanThrillerRating2                2.069e-01  8.328e-02   2.484 0.012983
## MeanThrillerRating3                2.476e-03  6.616e-02   0.037 0.970147
## MeanThrillerRating4                5.213e-02  6.542e-02   0.797 0.425593
## MeanThrillerRating5                1.313e-01  6.573e-02   1.997 0.045820
## War1                               6.323e-01  2.542e-02  24.880  < 2e-16
## MeanWarRating1                     1.357e-01  9.040e-02   1.502 0.133190
## MeanWarRating2                    -3.833e-02  4.375e-02  -0.876 0.381004
## MeanWarRating3                    -1.295e-01  2.604e-02  -4.974 6.56e-07
## MeanWarRating4                    -1.787e-01  2.366e-02  -7.551 4.33e-14
## MeanWarRating5                    -6.434e-02  2.395e-02  -2.686 0.007232
## Western1                           1.284e+00  5.896e-02  21.782  < 2e-16
## MeanWesternRating1                 2.891e-02  3.044e-02   0.950 0.342222
## MeanWesternRating2                 2.891e-02  1.971e-02   1.467 0.142458
## MeanWesternRating3                -1.640e-02  1.242e-02  -1.321 0.186569
```

```
## MeanWesternRating4                    1.525e-02  1.141e-02   1.336 0.181396
## MeanWesternRating5                    3.727e-02  1.264e-02   2.948 0.003195
## MeanMovieRating2                      1.015e+01  8.894e+01   0.114 0.909132
## MeanMovieRating3                      1.151e+01  8.894e+01   0.129 0.897023
## MeanMovieRating4                      1.297e+01  8.894e+01   0.146 0.884074
## MeanMovieRating5                      1.410e+01  8.894e+01   0.159 0.873999
## MeanUserRating3                       1.168e+00  2.527e-01   4.621 3.82e-06
## MeanUserRating4                       1.130e+00  2.531e-01   4.466 7.98e-06
## MeanUserRating5                       1.158e+00  2.536e-01   4.566 4.98e-06
## Action1:MeanActionRating1            -1.220e+01  7.653e+01  -0.159 0.873395
## Action1:MeanActionRating2            -1.432e+00  1.487e-01  -9.626  < 2e-16
## Action1:MeanActionRating3            -8.527e-01  3.024e-02 -28.198  < 2e-16
## Action1:MeanActionRating4            -5.111e-01  2.512e-02 -20.347  < 2e-16
## Action1:MeanActionRating5                    NA         NA      NA       NA
## Adventure1:MeanAdventureRating1      -1.229e+01  7.170e+01  -0.171 0.863935
## Adventure1:MeanAdventureRating2      -1.680e+00  1.918e-01  -8.760  < 2e-16
## Adventure1:MeanAdventureRating3      -7.480e-01  4.085e-02 -18.312  < 2e-16
## Adventure1:MeanAdventureRating4      -5.176e-01  3.542e-02 -14.616  < 2e-16
## Adventure1:MeanAdventureRating5              NA         NA      NA       NA
## Animation1:MeanAnimationRating1      -1.304e+01  4.026e+01  -0.324 0.746002
## Animation1:MeanAnimationRating2      -2.836e+00  2.456e-01 -11.550  < 2e-16
## Animation1:MeanAnimationRating3      -1.576e+00  6.356e-02 -24.801  < 2e-16
## Animation1:MeanAnimationRating4      -7.164e-01  4.304e-02 -16.646  < 2e-16
## Animation1:MeanAnimationRating5              NA         NA      NA       NA
## Children.s1:MeanChildren.sRating1    -1.157e+01  3.316e+01  -0.349 0.727201
## Children.s1:MeanChildren.sRating2    -1.941e+00  1.168e-01 -16.626  < 2e-16
## Children.s1:MeanChildren.sRating3    -1.133e+00  4.899e-02 -23.125  < 2e-16
## Children.s1:MeanChildren.sRating4    -6.142e-01  4.007e-02 -15.326  < 2e-16
## Children.s1:MeanChildren.sRating5            NA         NA      NA       NA
## Comedy1:MeanComedyRating1            -1.271e+01  8.671e+01  -0.147 0.883441
## Comedy1:MeanComedyRating2            -2.455e+00  2.510e-01  -9.782  < 2e-16
## Comedy1:MeanComedyRating3            -8.905e-01  2.538e-02 -35.088  < 2e-16
## Comedy1:MeanComedyRating4            -5.198e-01  1.946e-02 -26.710  < 2e-16
## Comedy1:MeanComedyRating5                    NA         NA      NA       NA
## Crime1:MeanCrimeRating1              -1.369e+01  7.338e+01  -0.187 0.852028
## Crime1:MeanCrimeRating2              -3.312e+00  3.823e-01  -8.665  < 2e-16
## Crime1:MeanCrimeRating3              -1.317e+00  4.735e-02 -27.827  < 2e-16
## Crime1:MeanCrimeRating4              -6.243e-01  3.045e-02 -20.505  < 2e-16
## Crime1:MeanCrimeRating5                      NA         NA      NA       NA
## Documentary1:MeanDocumentaryRating1 -1.593e+01  5.696e+01  -0.280 0.779690
## Documentary1:MeanDocumentaryRating2 -1.589e+01  5.412e+01  -0.294 0.769133
## Documentary1:MeanDocumentaryRating3 -3.961e+00  1.607e-01 -24.648  < 2e-16
## Documentary1:MeanDocumentaryRating4 -1.435e+00  9.766e-02 -14.694  < 2e-16
## Documentary1:MeanDocumentaryRating5          NA         NA      NA       NA
## Drama1:MeanDramaRating1              -1.682e+01  2.242e+02  -0.075 0.940196
## Drama1:MeanDramaRating2              -2.010e+00  2.023e-01  -9.933  < 2e-16
## Drama1:MeanDramaRating3              -1.116e+00  3.141e-02 -35.510  < 2e-16
## Drama1:MeanDramaRating4              -5.473e-01  1.478e-02 -37.040  < 2e-16
## Drama1:MeanDramaRating5                      NA         NA      NA       NA
## Fantasy1:MeanFantasyRating1          -1.307e+01  5.634e+01  -0.232 0.816517
```

```
## Fantasy1:MeanFantasyRating2          -2.828e+00  2.276e-01 -12.425  < 2e-16
## Fantasy1:MeanFantasyRating3          -1.738e+00  7.550e-02 -23.020  < 2e-16
## Fantasy1:MeanFantasyRating4          -9.496e-01  6.865e-02 -13.832  < 2e-16
## Fantasy1:MeanFantasyRating5                  NA         NA      NA       NA
## Film.Noir1:MeanFilm.NoirRating1      -1.421e+01  8.290e+01  -0.171 0.863948
## Film.Noir1:MeanFilm.NoirRating2      -5.418e+00  7.255e-01  -7.468 8.14e-14
## Film.Noir1:MeanFilm.NoirRating3      -3.121e+00  1.166e-01 -26.759  < 2e-16
## Film.Noir1:MeanFilm.NoirRating4      -1.036e+00  5.564e-02 -18.620  < 2e-16
## Film.Noir1:MeanFilm.NoirRating5              NA         NA      NA       NA
## Horror1:MeanHorrorRating1            -1.425e+01  4.627e+01  -0.308 0.758171
## Horror1:MeanHorrorRating2            -3.467e+00  1.242e-01 -27.926  < 2e-16
## Horror1:MeanHorrorRating3            -1.612e+00  4.869e-02 -33.119  < 2e-16
## Horror1:MeanHorrorRating4            -1.005e+00  4.537e-02 -22.138  < 2e-16
## Horror1:MeanHorrorRating5                    NA         NA      NA       NA
## Musical1:MeanMusicalRating1          -1.382e+01  4.891e+01  -0.283 0.777532
## Musical1:MeanMusicalRating2          -2.881e+00  1.918e-01 -15.015  < 2e-16
## Musical1:MeanMusicalRating3          -1.782e+00  5.767e-02 -30.904  < 2e-16
## Musical1:MeanMusicalRating4          -8.233e-01  4.077e-02 -20.195  < 2e-16
## Musical1:MeanMusicalRating5                  NA         NA      NA       NA
## Mystery1:MeanMysteryRating1          -1.319e+01  6.336e+01  -0.208 0.835105
## Mystery1:MeanMysteryRating2          -3.823e+00  3.548e-01 -10.776  < 2e-16
## Mystery1:MeanMysteryRating3          -1.623e+00  6.393e-02 -25.390  < 2e-16
## Mystery1:MeanMysteryRating4          -7.983e-01  4.737e-02 -16.850  < 2e-16
## Mystery1:MeanMysteryRating5                  NA         NA      NA       NA
## Romance1:MeanRomanceRating1          -1.403e+01  8.941e+01  -0.157 0.875333
## Romance1:MeanRomanceRating2          -2.522e+00  3.065e-01  -8.228  < 2e-16
## Romance1:MeanRomanceRating3          -1.163e+00  3.532e-02 -32.939  < 2e-16
## Romance1:MeanRomanceRating4          -6.125e-01  2.544e-02 -24.077  < 2e-16
## Romance1:MeanRomanceRating5                  NA         NA      NA       NA
## Sci.Fi1:MeanSci.FiRating1            -1.281e+01  7.214e+01  -0.178 0.859027
## Sci.Fi1:MeanSci.FiRating2            -2.630e+00  2.103e-01 -12.506  < 2e-16
## Sci.Fi1:MeanSci.FiRating3            -8.739e-01  3.797e-02 -23.014  < 2e-16
## Sci.Fi1:MeanSci.FiRating4            -5.303e-01  3.402e-02 -15.588  < 2e-16
## Sci.Fi1:MeanSci.FiRating5                    NA         NA      NA       NA
## Thriller1:MeanThrillerRating1        -1.455e+01  1.091e+02  -0.133 0.893922
## Thriller1:MeanThrillerRating2        -3.104e+00  3.041e-01 -10.206  < 2e-16
## Thriller1:MeanThrillerRating3        -9.255e-01  3.230e-02 -28.648  < 2e-16
## Thriller1:MeanThrillerRating4        -5.829e-01  2.372e-02 -24.577  < 2e-16
## Thriller1:MeanThrillerRating5                NA         NA      NA       NA
## War1:MeanWarRating1                  -1.406e+01  8.051e+01  -0.175 0.861413
## War1:MeanWarRating2                  -3.563e+00  3.143e-01 -11.336  < 2e-16
## War1:MeanWarRating3                  -1.397e+00  5.970e-02 -23.396  < 2e-16
## War1:MeanWarRating4                  -6.469e-01  2.949e-02 -21.933  < 2e-16
## War1:MeanWarRating5                          NA         NA      NA       NA
## Western1:MeanWesternRating1          -1.394e+01  4.978e+01  -0.280 0.779374
## Western1:MeanWesternRating2          -4.877e+00  3.523e-01 -13.841  < 2e-16
## Western1:MeanWesternRating3          -2.653e+00  8.395e-02 -31.596  < 2e-16
## Western1:MeanWesternRating4          -1.251e+00  6.396e-02 -19.558  < 2e-16
## Western1:MeanWesternRating5                  NA         NA      NA       NA
##
```

```
## (Intercept)
## GenderM                          **
## Age18-24                         *
## Age25-34
## Age35-44                         **
## Age45-49                         ***
## Age50-55                         ***
## Age56+                           .
## Occupationacademic/educator      *
## Occupationartist                 ***
## Occupationclerical/admin
## Occupationcollege/grad student   *
## Occupationcustomer service       **
## Occupationdoctor/health care     *
## Occupationexecutive/managerial
## Occupationfarmer                 ***
## Occupationhomemaker
## OccupationK-12 student           ***
## Occupationlawyer
## Occupationprogrammer             **
## Occupationretired
## Occupationsales/marketing        .
## Occupationscientist              **
## Occupationself-employed
## Occupationtechnician/engineer
## Occupationtradesman/craftsman    **
## Occupationunemployed             **
## Occupationwriter
## Action1                          ***
## MeanActionRating1
## MeanActionRating2
## MeanActionRating3
## MeanActionRating4
## MeanActionRating5
## Adventure1                       ***
## MeanAdventureRating1
## MeanAdventureRating2
## MeanAdventureRating3             .
## MeanAdventureRating4
## MeanAdventureRating5             .
## Animation1                       ***
## MeanAnimationRating1
## MeanAnimationRating2
## MeanAnimationRating3
## MeanAnimationRating4             ***
## MeanAnimationRating5             ***
## Children.s1                      ***
## MeanChildren.sRating1
## MeanChildren.sRating2            ***
## MeanChildren.sRating3
```

```
## MeanChildren.sRating4              .
## MeanChildren.sRating5
## Comedy1                            ***
## MeanComedyRating1
## MeanComedyRating2
## MeanComedyRating3
## MeanComedyRating4
## MeanComedyRating5
## Crime1                             ***
## MeanCrimeRating1                   *
## MeanCrimeRating2
## MeanCrimeRating3
## MeanCrimeRating4
## MeanCrimeRating5                   .
## Documentary1                       ***
## MeanDocumentaryRating1             .
## MeanDocumentaryRating2             *
## MeanDocumentaryRating3             ***
## MeanDocumentaryRating4             ***
## MeanDocumentaryRating5
## Drama1                             ***
## MeanDramaRating1                   ***
## MeanDramaRating2
## MeanDramaRating3
## MeanDramaRating4
## MeanDramaRating5
## Fantasy1                           ***
## MeanFantasyRating1
## MeanFantasyRating2                 ***
## MeanFantasyRating3                 **
## MeanFantasyRating4                 ***
## MeanFantasyRating5                 ***
## Film.Noir1                         ***
## MeanFilm.NoirRating1               **
## MeanFilm.NoirRating2               ***
## MeanFilm.NoirRating3               ***
## MeanFilm.NoirRating4               ***
## MeanFilm.NoirRating5               ***
## Horror1                            ***
## MeanHorrorRating1
## MeanHorrorRating2
## MeanHorrorRating3                  ***
## MeanHorrorRating4                  *
## MeanHorrorRating5                  .
## Musical1                           ***
## MeanMusicalRating1                 **
## MeanMusicalRating2                 ***
## MeanMusicalRating3                 *
## MeanMusicalRating4
## MeanMusicalRating5                 .
```

```
## Mystery1                          ***
## MeanMysteryRating1
## MeanMysteryRating2
## MeanMysteryRating3                *
## MeanMysteryRating4                ***
## MeanMysteryRating5                **
## Romance1                          ***
## MeanRomanceRating1                **
## MeanRomanceRating2
## MeanRomanceRating3
## MeanRomanceRating4
## MeanRomanceRating5
## Sci.Fi1                           ***
## MeanSci.FiRating1
## MeanSci.FiRating2
## MeanSci.FiRating3
## MeanSci.FiRating4
## MeanSci.FiRating5                 .
## Thriller1                         ***
## MeanThrillerRating1               **
## MeanThrillerRating2               *
## MeanThrillerRating3
## MeanThrillerRating4
## MeanThrillerRating5               *
## War1                              ***
## MeanWarRating1
## MeanWarRating2
## MeanWarRating3                    ***
## MeanWarRating4                    ***
## MeanWarRating5                    **
## Western1                          ***
## MeanWesternRating1
## MeanWesternRating2
## MeanWesternRating3
## MeanWesternRating4
## MeanWesternRating5                **
## MeanMovieRating2
## MeanMovieRating3
## MeanMovieRating4
## MeanMovieRating5
## MeanUserRating3                   ***
## MeanUserRating4                   ***
## MeanUserRating5                   ***
## Action1:MeanActionRating1
## Action1:MeanActionRating2         ***
## Action1:MeanActionRating3         ***
## Action1:MeanActionRating4         ***
## Action1:MeanActionRating5
## Adventure1:MeanAdventureRating1
## Adventure1:MeanAdventureRating2   ***
```

```
## Adventure1:MeanAdventureRating3     ***
## Adventure1:MeanAdventureRating4     ***
## Adventure1:MeanAdventureRating5
## Animation1:MeanAnimationRating1
## Animation1:MeanAnimationRating2     ***
## Animation1:MeanAnimationRating3     ***
## Animation1:MeanAnimationRating4     ***
## Animation1:MeanAnimationRating5
## Children.s1:MeanChildren.sRating1
## Children.s1:MeanChildren.sRating2   ***
## Children.s1:MeanChildren.sRating3   ***
## Children.s1:MeanChildren.sRating4   ***
## Children.s1:MeanChildren.sRating5
## Comedy1:MeanComedyRating1
## Comedy1:MeanComedyRating2           ***
## Comedy1:MeanComedyRating3           ***
## Comedy1:MeanComedyRating4           ***
## Comedy1:MeanComedyRating5
## Crime1:MeanCrimeRating1
## Crime1:MeanCrimeRating2             ***
## Crime1:MeanCrimeRating3             ***
## Crime1:MeanCrimeRating4             ***
## Crime1:MeanCrimeRating5
## Documentary1:MeanDocumentaryRating1
## Documentary1:MeanDocumentaryRating2
## Documentary1:MeanDocumentaryRating3 ***
## Documentary1:MeanDocumentaryRating4 ***
## Documentary1:MeanDocumentaryRating5
## Drama1:MeanDramaRating1
## Drama1:MeanDramaRating2             ***
## Drama1:MeanDramaRating3             ***
## Drama1:MeanDramaRating4             ***
## Drama1:MeanDramaRating5
## Fantasy1:MeanFantasyRating1
## Fantasy1:MeanFantasyRating2         ***
## Fantasy1:MeanFantasyRating3         ***
## Fantasy1:MeanFantasyRating4         ***
## Fantasy1:MeanFantasyRating5
## Film.Noir1:MeanFilm.NoirRating1
## Film.Noir1:MeanFilm.NoirRating2     ***
## Film.Noir1:MeanFilm.NoirRating3     ***
## Film.Noir1:MeanFilm.NoirRating4     ***
## Film.Noir1:MeanFilm.NoirRating5
## Horror1:MeanHorrorRating1
## Horror1:MeanHorrorRating2           ***
## Horror1:MeanHorrorRating3           ***
## Horror1:MeanHorrorRating4           ***
## Horror1:MeanHorrorRating5
## Musical1:MeanMusicalRating1
## Musical1:MeanMusicalRating2         ***
```

```
## Musical1:MeanMusicalRating3         ***
## Musical1:MeanMusicalRating4         ***
## Musical1:MeanMusicalRating5
## Mystery1:MeanMysteryRating1
## Mystery1:MeanMysteryRating2         ***
## Mystery1:MeanMysteryRating3         ***
## Mystery1:MeanMysteryRating4         ***
## Mystery1:MeanMysteryRating5
## Romance1:MeanRomanceRating1
## Romance1:MeanRomanceRating2         ***
## Romance1:MeanRomanceRating3         ***
## Romance1:MeanRomanceRating4         ***
## Romance1:MeanRomanceRating5
## Sci.Fi1:MeanSci.FiRating1
## Sci.Fi1:MeanSci.FiRating2           ***
## Sci.Fi1:MeanSci.FiRating3           ***
## Sci.Fi1:MeanSci.FiRating4           ***
## Sci.Fi1:MeanSci.FiRating5
## Thriller1:MeanThrillerRating1
## Thriller1:MeanThrillerRating2       ***
## Thriller1:MeanThrillerRating3       ***
## Thriller1:MeanThrillerRating4       ***
## Thriller1:MeanThrillerRating5
## War1:MeanWarRating1
## War1:MeanWarRating2                 ***
## War1:MeanWarRating3                 ***
## War1:MeanWarRating4                 ***
## War1:MeanWarRating5
## Western1:MeanWesternRating1
## Western1:MeanWesternRating2         ***
## Western1:MeanWesternRating3         ***
## Western1:MeanWesternRating4         ***
## Western1:MeanWesternRating5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 954859  on 700148  degrees of freedom
## Residual deviance: 770490  on 699934  degrees of freedom
## AIC: 770920
##
## Number of Fisher Scoring iterations: 12
```

Random Forest:

For Random Forest, the Out-Of-Bag error rate, is 31.03%, which is fairly low error rate. This means that the model is consistent over most of the trees created. When looking at variable importance, we can see that Random Forest placed a much higher importance on MeanMovieRating than the GLM did. It was b

y far the most important predictor in the model. After this, Occupation, Age and MeanUser Rating were important user-specific factors, along with the mean user ratings for a variety of genres, but for sci-fi and comedy movies most o f all. The genre of the movie was important as well, particular whether or no t it was a comedy, drama, thriller or action movie.

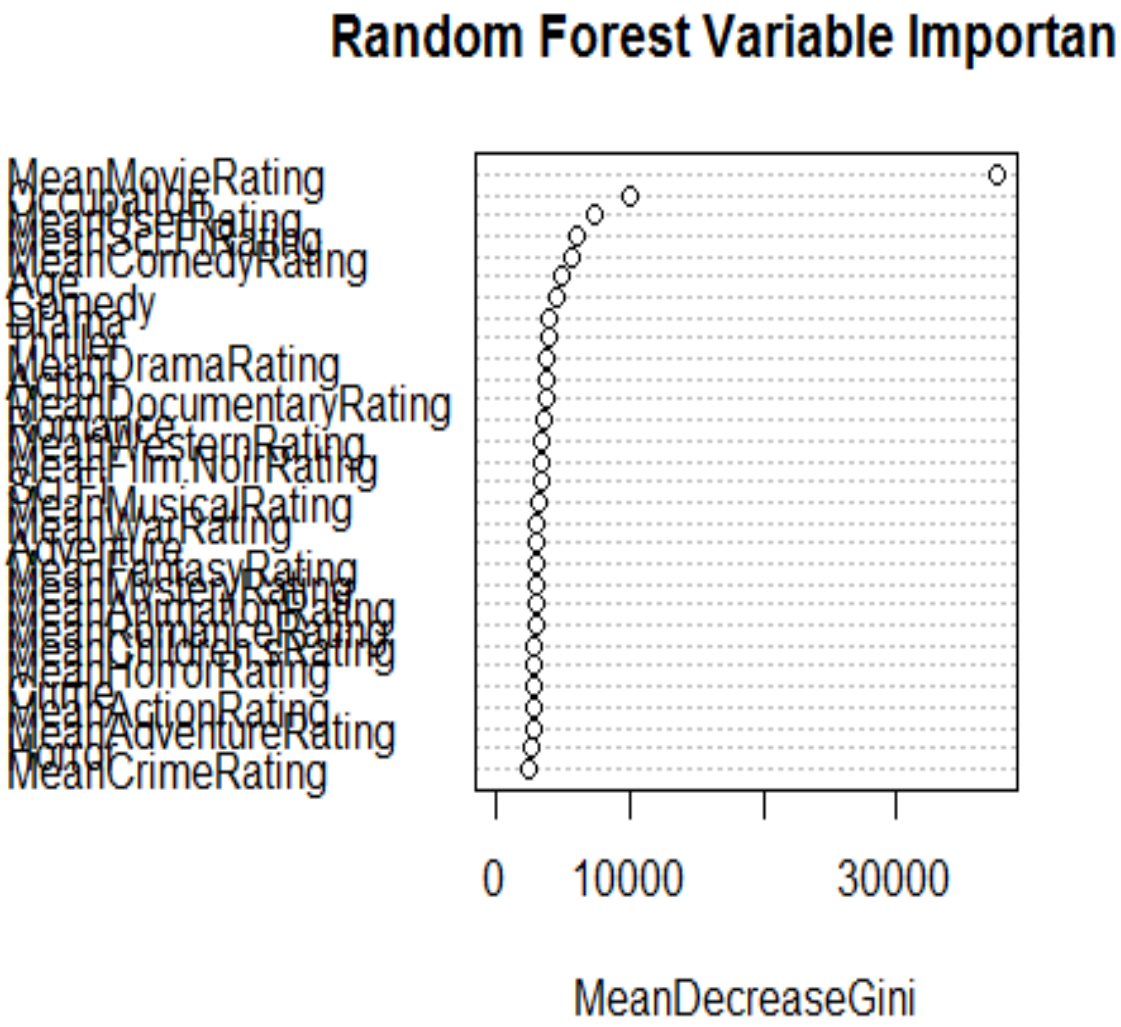
The interaction variables did not have the powerful effects that they did in the GLM model. This is likely because the trees of the forest were able to replicate the point of an interaction variable by making two "branches out of the genre and mean genre rating variables.

The sensitivity is 79.47% and the specificity is 58.6%. The positive pre dictive power is 72.27% and the negative predictive power is 67.77%, so the overall accuracy is 70.62%. As a predictive model, Random Forest is objective ly better than GLM at predicting results, even though it has lost some of its positive predictive power and specificity. It has made massive gains in terms of accuracy, negative predictive power and sensitivity.

```
#Random Forest Results
confusionMatrix(forestpred, test$Rating, positive="Like")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Dislike   Like
##    Dislike   74573  35471
##    Like      52693 137323
##
##                Accuracy : 0.7062
##                  95% CI : (0.7045, 0.7078)
##     No Information Rate : 0.5759
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3876
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.7947
##             Specificity : 0.5860
##          Pos Pred Value : 0.7227
##          Neg Pred Value : 0.6777
##              Prevalence : 0.5759
##          Detection Rate : 0.4577
##    Detection Prevalence : 0.6333
##       Balanced Accuracy : 0.6903
##
##        'Positive' Class : Like
##
```

```
varImpPlot(fit.forest, type=2, main=" Random Forest Variable Importance")
```



Boosting:

The results were almost identical to those of random forest. The sensitivity is 82.98% and the specificity is 52.37%. The positive predictive power is 70.28% and the negative predictive power is 69.38%, for an overall accuracy of 70%. This model was also very good at predicting likes. There were very few false positives. It was much worse at figuring what movies users would dislike, so there are many false negatives.

The most important predictor of the results was by far MeanMovieRating. Demographic factors did not matter at all. The genre of the movie did not matter at all either. The users' mean average ratings for movie genres and average rating for all movies did matter, but the importance for many of these factors was small. The mean drama rating was the most important factor after MeanMovieRating, trailed by MeanUserRating. All other factors had little to no importance at all. This is likely because we limited our model to building cla

```
ssifiers of size 5 or less. This means that many of the more minor factors co
uld not be included. This is a very simple model compared to the past two bec
ause it has so few predictors.
```

```
# Boosting Results
confusionMatrix(boostpred$class, test$Rating, positive="Like")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Dislike   Like
##    Dislike   66643  29409
##    Like      60623 143385
##
##                Accuracy : 0.7
##                  95% CI : (0.6983, 0.7016)
##     No Information Rate : 0.5759
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.3653
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.8298
##             Specificity : 0.5237
##          Pos Pred Value : 0.7028
##          Neg Pred Value : 0.6938
##              Prevalence : 0.5759
##          Detection Rate : 0.4779
##    Detection Prevalence : 0.6799
##       Balanced Accuracy : 0.6767
##
##        'Positive' Class : Like
##
```

```
fit.boost$importance
```

```
##               Action            Adventure                  Age
##           0.02759853            0.00000000            0.00000000
##            Animation            Children.s               Comedy
##           0.00000000            0.00000000            0.00000000
##                Crime           Documentary                Drama
##           0.00000000            0.00000000            0.09109497
##              Fantasy             Film.Noir               Gender
##           0.00000000            0.00000000            0.00000000
##               Horror      MeanActionRating   MeanAdventureRating
##           0.00000000            0.11930488            0.40123448
##   MeanAnimationRating  MeanChildren.sRating      MeanComedyRating
##           0.10351111            0.00000000            5.66975268
##       MeanCrimeRating MeanDocumentaryRating       MeanDramaRating
##           0.04673609            0.01249720           16.49474441
##     MeanFantasyRating   MeanFilm.NoirRating      MeanHorrorRating
```

```
##            0.02202529            0.31602801            0.11383041
##       MeanMovieRating     MeanMusicalRating     MeanMysteryRating
##           67.03582124            0.05234272            0.02556014
##     MeanRomanceRating      MeanSci.FiRating    MeanThrillerRating
##            0.09377798            1.40541111            0.46690515
##        MeanUserRating         MeanWarRating     MeanWesternRating
##            6.14946752            1.35235606            0.00000000
##               Musical               Mystery            Occupation
##            0.00000000            0.00000000            0.00000000
##               Romance                Sci.Fi              Thriller
##            0.00000000            0.00000000            0.00000000
##                   War               Western
##            0.00000000            0.00000000
```

==========

DISCUSSION

GLM:

One of the most obvious flaws with the GLM model is that there are too many variables. The results are difficult to sift through. A simpler model may be preferred, merely for the sake of more clearly understanding how it functions. However, there is another slight problem with the GLM model. It predicted that users would dislike almost all movies. This is a problem because we want to be able to distinguish between correct and incorrect matches for users and movies. There is also a chance that a user will not be given a recommendation at all because the model is so stingy about predicting that a user will like a given movie. However, when faced with a choice between being able to predict what movies people will like and being able to predict what movies people won't like, we would strongly prefer the former. After all, nobody comes to a movie recommendation service in order to figure out which movies they would hate. What they want is a system that will only recommend movies that they will like (positive predictive power) and will never recommend movies that they will dislike (specificity). If a user disliked a movie, there is almost no way that our model would recommend it to them. If our model recommended a movie, it is highly likely that the user will enjoy it. In this sense, our model is a success! Additionally, the small number of recommended movies are recommendations that we can be very sure that the users will enjoy because the model was so strict.

So the model is moving in the right direction, but we do need to find a way to increase its sensitivity and negative predictive power. There are thousands of movies that a user might miss out on because our model allows too many false negatives. Only 4% of hypothetically successful recommendations will actually get made. We should also keep in mind that there are 6040 users and the model made 7109 recommendations, so each user would only get one movie recommendation on average. If the model made a random prediction about any movie and any user, there is only a 32% chance that this prediction will be accurate. We should try to correct these faults, even if there is a slightly greater chance mismatching users and movies.

Random Forest:

Random Forest is a perfect technique to improve our model. Random Forest was far more likely than not to predict that a user would give a movie a positive review, the converse of GLM. After running it, we had a greatly improved sensitivity of 79.4% and negative predictive power of 67.77%. This is still not perfect, but it is vastly better than the GLM model. We do not need a perfect sensitivity, since the user does not need to see every movie that they might enjoy. They only need a handful to choose from. Users can now enjoy almost 80% the movies that would make good matches for them, which is likely more than enough. This fixes the most glaring problem with the GLM model.

The accuracy for the entire model is 70.62%. In other words, if I asked this model whether or not I would like a given movie, there's a 70% chance that it would be accurate. As an outcome-prediction system, that is not bad and definitely represents an improvement over GLM. As a movie recommendation system, it is effective but not as immaculate as GLM. The positive predictive power is 72.27%, so if the model predicted that I would like a movie, it would be correct roughly three-quarters of the time. This is a decrease from GLM, but still within acceptable bounds. The specificity is only 58.59%, so if a user would dislike a movie, there is still a 41.4% chance that our model would recommend it anyway. Our system has gotten much better at identifying "good" matches, but slightly worse at identifying "bad" matches.

The outcome is that users will have a far greater number of movies to choose from. However, they will still have to exercise some human intelligence when they choose, as about a quarter of the movies that our system recommends are false positives. This is not a bad thing. It will give users some agency over their own results. After all, this is how real movie recommendation systems work today. There is a tradeoff between the quantity of movies recommended and the quality of these recommendations. This model provides a reasonable balance between the two, since users can enjoy 79.4% of the movies that they should match with and recommendations will be accurate nearly 72.27% of the time.

One major downside of Random Forest is that it is a black box technique. There is no way to know where the variables went within the tree and what impact they made without decomposing the random forest into its individual trees and examining all of them. We can see which variables mattered the mattered, but not how they mattered.

One way to improve this model would be to use more trees. We used only 10 trees because we have so many variables. It would be difficult to run the program with a large number of trees, but this would be possible with more advanced hardware. Perhaps with more trees the model would become better at identifying dislikes and would obtain a higher specificity.

Boosting:

Like Random Forest, Boosting is a powerful black box technique. It could also improve on GLM's flaws. With boosting, there is no need to include interaction variables because the boosting function can create interactions internally. As with Random Forest, we faced hardware limitations. With so many variables, the boosting function could have benefitted from having very "deep" cl

assifiers. Due to hardware considerations, we had to limit our strong classifiers to a maximum depth of 5. This makes the model a very simple one and the result is that most of the predictors did not make it into the model.

This model's similarity to the Random Forest model should inspire some confidence, especially since it is so simple. It is reassuring to see that the Boosting function also selected MeanMovieRatings, MeanUserRating and the mean genre ratings as the most important variables, even though these are the obvious candidates for predicting whether or not a user will like a movie. It is somewhat troubling that it was never able to include any of the other variables. Perhaps the model was too simple.

Like Random Forest, this model is a fairly accurate predictor, but an imperfect recommender. The two functions share almost identical outcomes. The accuracy, both predictive powers and specificity have decreased slightly but the sensitivity has risen. This means that our model is slightly better at identifying good matches. It is worse at identifying worse matches. With a specificity of 52.36%, it is barely better than a guess. This is a tradeoff that, in our opinion, does not pay off. The random forest results delivered an adequate number of true positives. Its main fault was its poor specificity, and on this front Boosting has unfortunately failed to improve the results. This is ironic since the boosting technique is built for classification.

This does shed some additional light on how to improve the Random Forest model. The Random Forest model was better because it was able to take advantage of demographic and movie specific factors. This is what helped it improve its specificity, which was the major problem that it faced and that boosting only made worse. One of the best ways to improve the Random Forest technique might be to include more demographic and movie specific factors. If there is a way to describe movies beyond their genre, perhaps by release year or by their descriptions, this could be very effective. There are always more demographic factors that we could obtain about users, including their location, race, religious views, and more.

==========

CONCLUSION

Ultimately, it seems that random forest provided the best model in terms of both predictive power and as a recommendation system. The best way to continue improving the model is likely to include more trees in the model and to obtain more information about our movies and users. The best version of the model gives predictions that are accurate about 72% of the time and is able to recommend 80% of the movies that a user would hypothetically enjoy. Ultimately, there may be a tradeoff between the two numbers, and the right outcome is subjective. We believe that this is an optimal tradeoff and a good recommendation model.

==========

REFERENCES:

F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History
and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages. DOI=http://dx.doi.org/10.1145/2827872