Business Intelligence: Analysis of MovieLen Data

DS 501: Case Study 2 Report

Hang Ding, Fangling Zhang, Qingquan Zhao, Yihao Zhou, Tongge Zhu

1. Introduction

Business intelligence (BI) is a technology-driven process for analyzing data and presenting actionable information to help corporate executives, business managers and other end users make more informed business decisions. In this BI report, "Movielens 1M" Data set is analyzed to explore people's opinions on movies. The Data files obtained from http://grouplens.org/datasets/movielens/ contain 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users who joined MovieLens in 2000. It is spread across 3 tables: "movies.dat", "rating.dat" and "users.dat". The "users.dat" file contains user information from the following aspects: UserID, Gender, Age, Occupation and Zip Code (locations). The "movies.dat" file describes movie information about Movie ID, title and Genres. The total number of genre type is 18, such as Action, Musical etc, and each movie has at least one genre. The "rating.dat" file provides information about UserID, MovieID, Rating and Timestamp. We combined these three different tables by using the overlapping names as the merge (or join) keys, and saved it as an HDF5 file (shown in CaseStudy2_Team2.ipynb Out [1], Out [135]).

In this study case, by using Python libraries including Pandas, Matplotlib, Numpy and Scikit-learn, we interpreted and analyzed the data values and connections among users information, movies information and rating information. In addition, the following two interesting business intelligence question are answered:
(1) If we have a new movie, what should we do for best advertising?
(2) If we have a new movie, and we purchased some information about how many males/females watched it (e.g. from YouTube), and what occupations and ages these people are (e.g. from Internet provider company), but we do not know if people like it or not. Besides box-office information in theatre, can we get a rough idea if people who watched this movie like it or not?

2. Basic detail of MovieLen 1M Data set

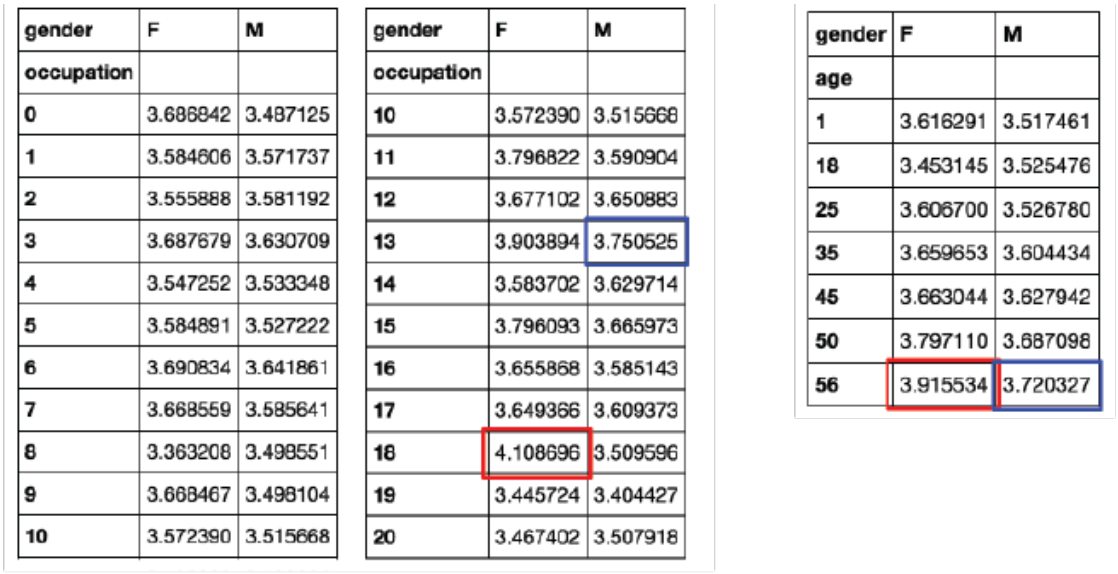
As the starting point of the analysis, a series of basic rating information were first summarized. The results are described as follows. Only 29 movies had an average score over 4.5 stars (shown in CaseStudy2_Team2.ipynb Out [3]) among over one million ratings. 29 movies have an average score over 4.5 among men, whereas 70 movies above 4.5 were obtained rated by women (shown in CaseStudy2_Team2.ipynb Out [4][5]). Interestingly, these two numbers increase to 105 and 187 respectively when we set an extra limit with age over 30 (shown in CaseStudy2_Team2.ipynb Out [4][5]).

We then grouped the data by movie title in order to obtain the rating times for each title, top ten of which are shown in Figure 1. We assume these ten movies are most popular movies because the number of ratings, which indicates the numbers of people who actually pay attentions to this movie, could reveal a movie's popularity more than other factors.

Title	# of rating
American Beauty (1999)	3428
Star Wars: Episode IV - A New Hope (1977)	2991
Star Wars: Episode V - The Empire Strikes Back (1980)	2990
Star Wars: Episode VI - Return of the Jedi (1983)	2883
Jurassic Park (1993)	2672
Saving Private Ryan (1998)	2653
Terminator 2: Judgment Day (1991)	2649
Matrix, The (1999)	2590
Back to the Future (1985)	2583
Silence of the Lambs, The (1991)	2578

Figure 1. Top Ten Most Popular Movies

To answer the question of what kinds of people are easier to be pleased, we grouped people by occupations (left panel of Figure 2) and by ages (right panel of Figure 2) versus genders. From the figure, we can see that over 56 years old female and female whose occupation is "tradesman/craftsman" gave an average score of 3.95 and 4.11, respectively, which are significant higher than score given by other groups, suggesting these two groups of people are easier to be pleased.



gender	F	M
occupation		
0	3.686842	3.487125
1	3.584606	3.571737
2	3.555888	3.581192
3	3.687679	3.630709
4	3.547252	3.533348
5	3.584891	3.527222
6	3.690834	3.641861
7	3.668559	3.585641
8	3.363208	3.498551
9	3.668467	3.498104
10	3.572390	3.515668

gender	F	M
occupation		
10	3.572390	3.515668
11	3.796822	3.590904
12	3.677102	3.650883
13	3.903894	3.750525
14	3.583702	3.629714
15	3.796093	3.665973
16	3.655868	3.585143
17	3.649366	3.609373
18	4.108696	3.509596
19	3.445724	3.404427
20	3.467402	3.507918

gender	F	M
age		
1	3.616291	3.517461
18	3.453145	3.525476
25	3.606700	3.526780
35	3.659653	3.604434
45	3.663044	3.627942
50	3.797110	3.687098
56	3.915534	3.720327

Figure 2. The data is grouped by occupations (left panel) and by ages (right panel) to analyze different genders' average scores.

2. **Expand our investigation to histograms**

To get more understandings of data, several histogram graphs were further plotted. The first one is the rating count of all movies (Figure 3), from where we see that most movies falls into the ratings among 3.0 to 5.0. The second graph we plotted is the number of ratings each movie received (Figure 4). The plot clearly shows an exponential relationship between number of movies and number of ratings, thus suggesting only a small portion of movies received a large amount of ratings, and most movies received less than 500 rating times.

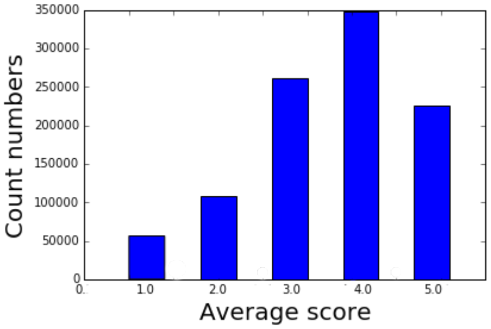


Figure 3. Histogram of rating count of all movies

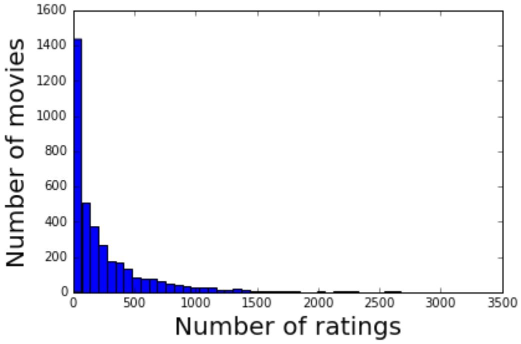


Figure 4. Histogram of number of ratings each movie received

Furthermore, two histograms of average rating of each movie (Figure 5) and average rating of each movie which had been rated for more than 100 times (Figure 6) are also generated. Both graphs show that the average ratings for most movies are in between 3 and 4. However, the comparison between these two figures indicate that the “tails” in low ratings and high ratings in Figure 5 disappear in histogram using the movies rated more 100 times (Figure 6). We would trust more on the high score of those movies being rated more 100 times as a good movie, instead of relying on those movies with very few rates. The reason for is that maybe the high score of those movies with only a very few rating numbers are only favored by a small number of people.

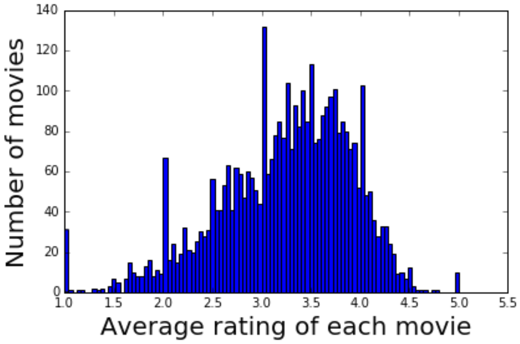


Figure 5. Average ratings of each movie

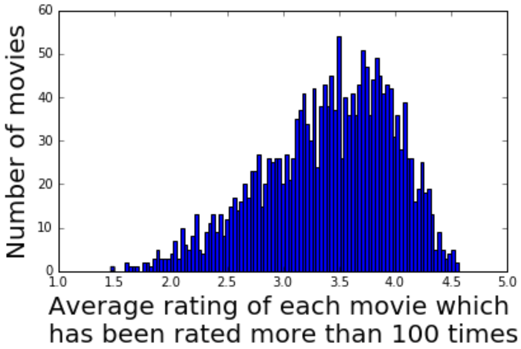
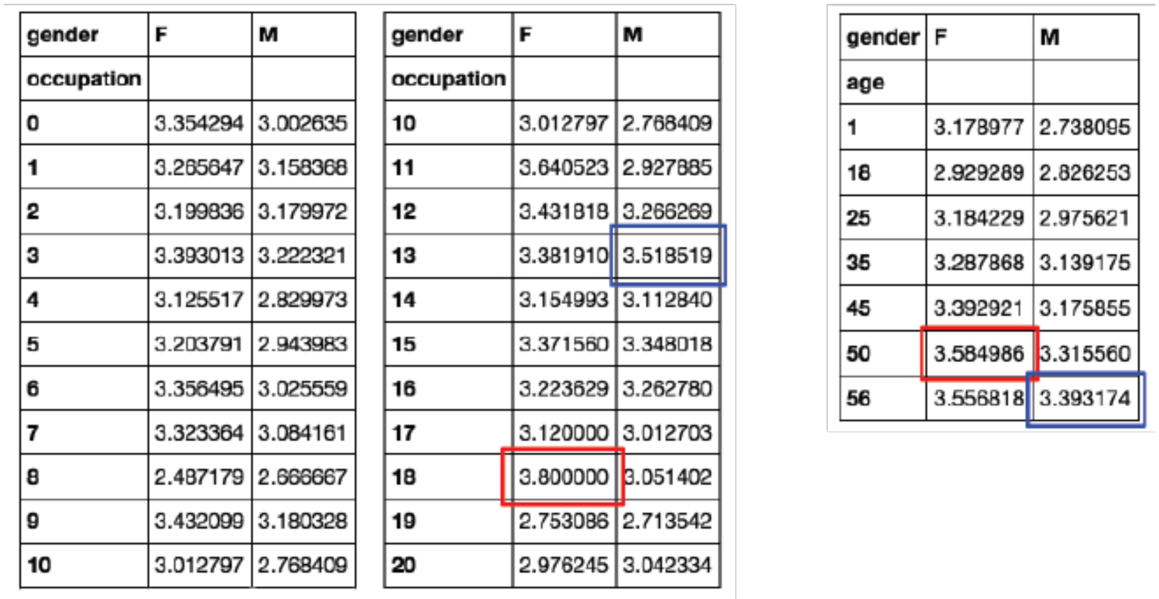


Figure 6. Average ratings of each movie being rated more than 100 times

We then used the movies being rated more than 100 times to plot the same idea as Figure 2, but more reliable this time as shown in Figure 7. Similarly as Figure 2, here we also found that tradesman/craftsman (3.8, shown in red box) are the easiest to be pleased as female and retired men (3.51, shown in blue box) are the easiest to be pleased as male. For age group, females in age group 50 (3.58) and males in age group 56 (3.39) are top easiest to be pleased. Both male and female in age 18 are the hardest to please. We think it is consistent with our conjectures because young people are more critics about the movies.



gender	F	M
occupation		
0	3.354294	3.002635
1	3.265647	3.158368
2	3.199836	3.179972
3	3.393013	3.222321
4	3.125517	2.829973
5	3.203791	2.943983
6	3.356495	3.025559
7	3.323364	3.084161
8	2.487179	2.666667
9	3.432099	3.180328
10	3.012797	2.768409

gender	F	M
occupation		
10	3.012797	2.768409
11	3.640523	2.927885
12	3.431818	3.266269
13	3.381910	3.518519
14	3.154993	3.112840
15	3.371560	3.348018
16	3.223629	3.262780
17	3.120000	3.012703
18	3.800000	3.051402
19	2.753086	2.713542
20	2.976245	3.042334

gender	F	M
age		
1	3.178977	2.738095
18	2.929289	2.826253
25	3.184229	2.975621
35	3.287868	3.139175
45	3.392921	3.175855
50	3.584986	3.315560
56	3.556818	3.393174

Figure 7. The data is grouped by occupations (left panel) and by ages (right panel) to analyze different genders' average scores.

3. Correlation: Men versus women¶

Next, we are interested in exploring gender different behaviors in scoring. Two scatter graphs are plotted to show men versus women and their mean rating for each movie (Figure 8 for all movies, and Figure 9 for those movies being rated more than 200 times.).

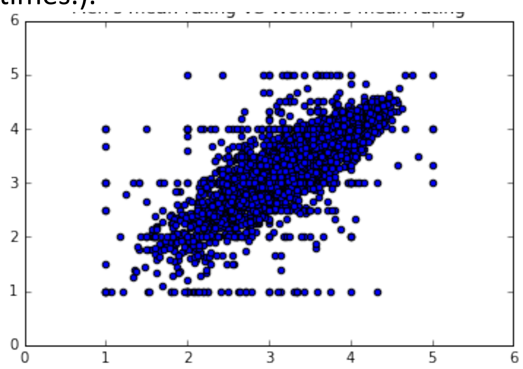


Figure 8. Men's mean rating VS Women's mean rating

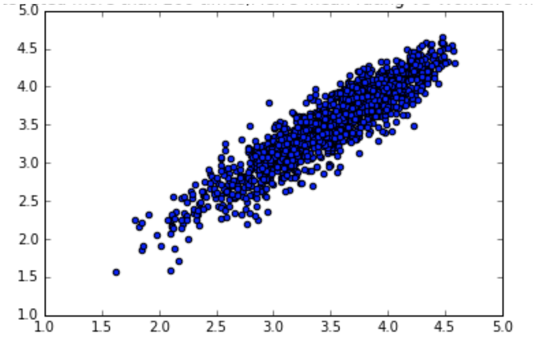


Figure 9. For movies rated more that 200 times, Men's mean rating VS Women's mean rating

From figure 8, we can see that Men's mean rating is correlated to Women's mean rating. By computing the data, we found the correlation coefficient number between Men's and Women's mean rating was 0.76 (shown in CaseStudy2_Team2.ipynb Out [22]), which support our observation. The comparing of Figure 8 and Figure 9 clearly shows that movies rated more than 200 times are more correlated between man and women's average ratings. A greater number of correlation coefficient (0.92 (shown in CaseStudy2_Team2.ipynb Out [23])) calculated for Figure 9 compared to the number from Figure 8 supported our observation.

Age	Correlation of Genders
1	0.3479
18	0.5756
25	0.6863
35	0.5994
45	0.5690
50	0.5369
56	0.4131

Figure 10. Correlation coefficient between the ratings of men and women for every age range

Occupation	Correlation of Occupation	Occupation	Correlation of Occupation
0	0.5788	11	0.3941
1	0.6364	12	0.4501
2	0.4724	13	0.2943
3	0.4388	14	0.5335
4	0.5726	15	0.4796
5	0.3298	16	0.4688
6	0.5185	17	0.5794
7	0.5727	18	0.2768
8	0.2752	19	0.4081
9	0.2766	20	0.6068
10	0.3305		

Figure 11.The correlation coefficient between the ratings of men and women for each occupation.

Next, We computed the correlation coefficient between the ratings of men and women for every age range (Figure 10). The largest correlation (0.69) is from people in age 25(age between 25-34). We can draw a conclusion here that for people between age 25 and 34, the rating given by one gender can be used to predict the rating given by the other gender.

Similarly, we also computed the correlation coefficient between the ratings of men and women for each occupation (Figure 11). We found that the largest correlation (0.64) is from people in occupation 1 (academic /educator). We can also conclude that for academic/educator, the rating given by one gender can be used to predict the rating given by the other gender.

4. Business Questions:

After being conducting several analyses in previous study, we are very interested in two important business questions:

(1) If we have a new movie, what should we do for best advertising?

(2) If we have a new movie, and we purchased some information about how many males/females watched it (e.g. from YouTube), and what occupations and ages these people are (e.g. from Internet provider company), but we do not know if people like it or not. Besides box-office information in theatre, can we get a rough idea if people who watched this movie like it or not?

4.1 Explore the first question:

To get the first answer, we have to first explore the relationships between a movie's genre type and customers' behaviors because for a new movie, we already know what genres this movie belongs to. Unfortunately, the 18 types of genres in the raw data we collected are mixed together for one movie; for example, one movie could belong to two or three types of genres. Therefore, we first slice each movie's genre data and map the information to 18 new columns with labeling "1" if this movie belongs to one genre and '0" otherwise.

With the information of whether a movie belongs to a single genre in hand, we first examined the numbers of movies of each genre as shown in Figure 12. Although duplicated counting numbers exist in the histogram, we can have a rough understanding that Comedy, Drama and Action are under much more focus compared with Documentary and Film-Noir movies.

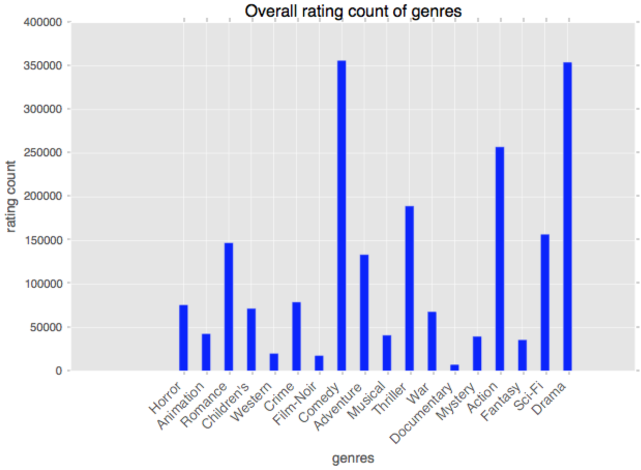


Figure 12. Rating numbers of each genre

4.1.1 Male versus female

We first plotted two histograms, male/female average ratings towards genres, which showed not much difference with a high coefficient of determination (Figure 13), and male/female rating numbers towards genres, which resulted that more male raters are involved in each genre (Figures not shown in this documents). A conclusion might be able to drawn from our data and analysis that more focus in terms of advertising should be put on male customers for a new movie of any genres. However, this conclusion is not solid enough because it cannot reflect male and female's different opinions on different genres since the overall male raters are much more than female raters.

Therefore, a more detailed histogram is plotted as shown in Figure 14 describing within male or female raters, what are the percentages of them who are interested in each genres.

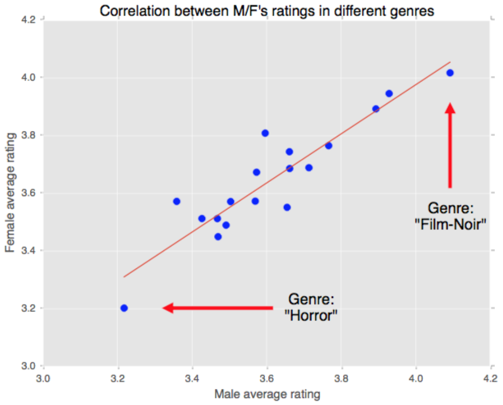


Figure 13. Coefficient of determination between male/female average ratings towards genres

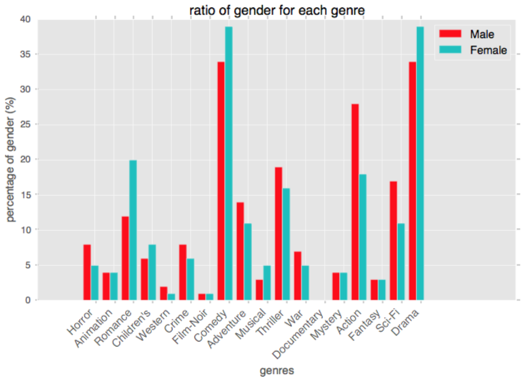


Figure 14. Percentage of genders that are interested in each genres

The y-axis of red bar represents the percentage of women raters who rate one genres in overall women raters, and the blue bar means the same but for male raters. Apparent differences are shown here that over 20% of female raters are rating Romance genre movies, whereas this number for male is only 12%. In contrast, about 28% of male raters are interested in talking about Action movies, but only 18% females do the same thing. So a conclusion can be made here that if the new movie is on a romantic topic, probably focusing on women customers and letting more women know this movie is a better advertising decision because higher percentage of females are more inclined in discuss about this movie. The situation is reversed if the new movie is about Action since male customers are easier to be attracted.

4.1.2 Different occupation's behaviors

After obtaining male/female opinions on genres, the occupation's effects on different genre were also explored. A series of histograms have been plotted to show how many ratings from each occupation were collected in data. A representative example towards Action genre is shown in Figure 15. All these series are not so useful and no solid conclusion can be drawn from them. For example, much fewer farmers are rating Action movies compared with other occupations, but the overall famer raters are also a very small number compared to others.

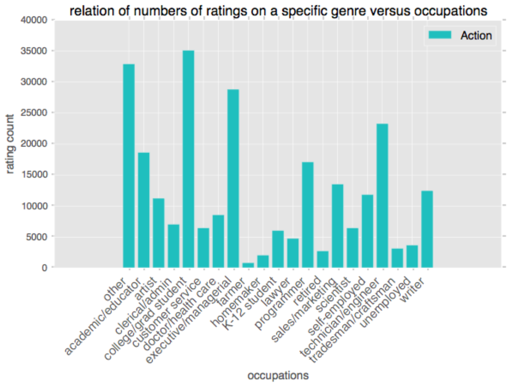


Figure 15. Rating counting of each occupation in Action movies

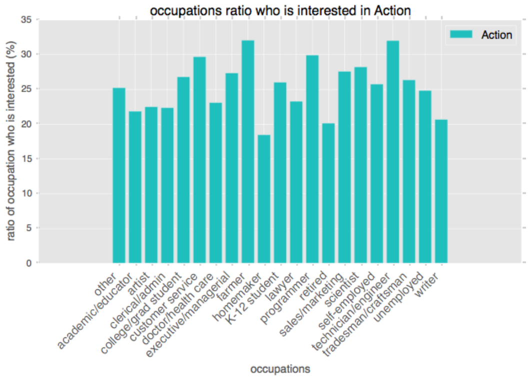


Figure 16. Percentage in each occupation interested in Action movie

Therefore, a series of modified graphs were prepared to describe the percentage of people in an occupation who rated Action movie within this occupation's overall raters. As the same example towards Action movie as shown in Figure 16, over 30% of farmer raters are talking about Action movies, even more than college/graduate student, who has the largest number in the previous figure. From this, we have to say if our new movie is an action movie, we probably need to take both factors into considerations.

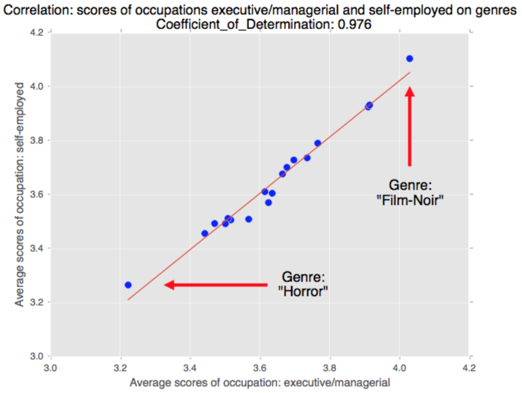


Figure 17. "Executive/managerial" and "self-employed" opinions on all genres. Each dot represents average score of a genre.

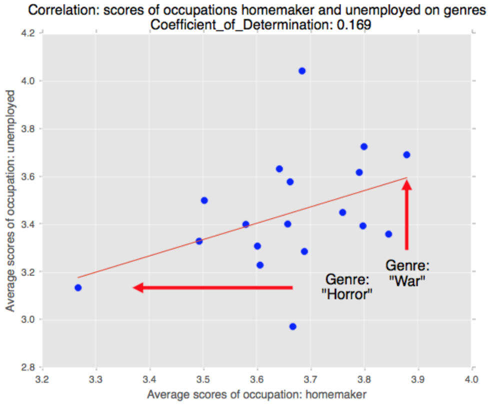


Figure 18. "Homemaker" and "unemployed" opinions on all genres. Each dot represents average score of a genre.

Beside the above analysis on a specific genre, we also want to know if any two occupations have similar opinions on all genres. Therefore, we calculated the coefficient of determination on average scores of all combinations of two occupations toward genres. One of the top similar behaviors comes from "executive/managerial" and "self-employed" with a coefficient of determination of 0.976 is shown in Figure 17. In contrast, "homemaker" and "unemployed" showed complete behaviors on almost all genres with a coefficient of determination of 0.19 (Figure 18). Although these are just two extreme examples, they clearly indicate that for advertising purpose, we can combine "executive/managerial" and "self-employed" occupations together as the same target, whereas we have to focus separately for "homemaker" and "unemployed".

To sum up this section, we could conclude that in order to better advertising a new movie, we should first look at if different occupations have similar opinions on the movie's genres. If so, we can pack them up as one group for one advertising strategy. Also, we have to consider both the numbers of people and the ratio of people in one occupation who are interested in this genre.

4.1.3 Locations. Figure 19 answers the question where we should focus on more for advertising simply because there are more people involved in rating. Although there might be many reasons such as a large population or more active people who are interested in rating, this data clearly indicates that California, New York, Minnesota, Illinois and Texas are several state with highest priorities compared to Mississippi and North Dakota.

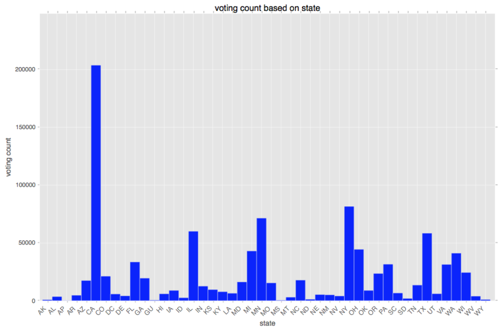


Figure 19. Rating count based on states

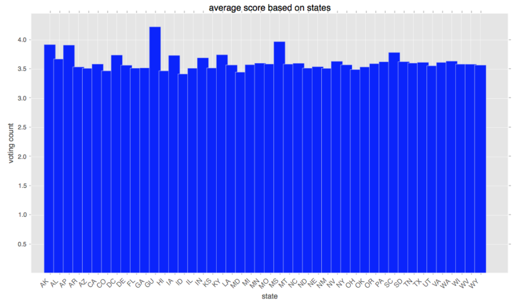


Figure 20. Average rating score of states

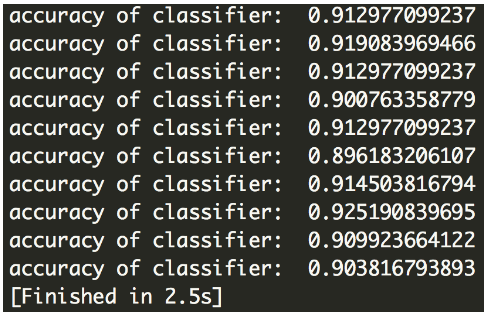
Another perspective for choosing location is by looking at the average score of each state (Figure 20). However, not many differences are obtained except Guam, which however has very few raters. Therefore, people from different states have similar opinions and we probably should not put too much thought on this.

4.2 Explore the second question:

To get the second answer, we did a supervised machine learning on each movie's features, and thus can predict if a new movie that has these features falls into different class of ratings. Luckily, we obtained an around 90% accuracy classifier to do this.

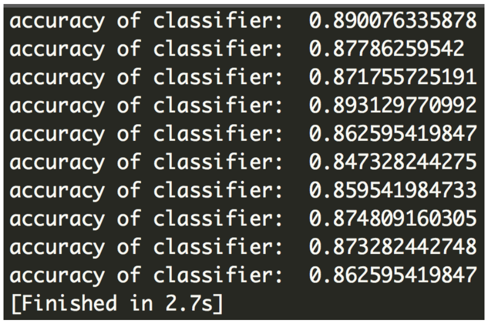
The procedure is described in the following. We first setup a new data frame with index of each movie. The features include two columns describing male rating counts and female rating counts, 21 columns of occupation's counts from occupation 0 to 20, 7 columns of different age's rating counts, and finally, 18 columns of genres information, each of which contains either 1 for this movie belongs to this genre or 0 indicating this movie is this genre.

For classification of each movie, we tried to map all movies average scores to either above 4.0, indicating a great movie, or below 4.0, which means it is not a great movie. In this way, by using either K-NN or SVM, we obtained an around 90% accuracy in prediction (Figure 21).



```
accuracy of classifier:  0.912977099237
accuracy of classifier:  0.919083969466
accuracy of classifier:  0.912977099237
accuracy of classifier:  0.900763358779
accuracy of classifier:  0.912977099237
accuracy of classifier:  0.896183206107
accuracy of classifier:  0.914503816794
accuracy of classifier:  0.925190839695
accuracy of classifier:  0.909923664122
accuracy of classifier:  0.903816793893
[Finished in 2.5s]
```

Figure 21. Supervised machine learning accuracy with 2 classes (below 4.0 or above 4.0), running 10 times



```
accuracy of classifier:  0.890076335878
accuracy of classifier:  0.87786259542
accuracy of classifier:  0.871755725191
accuracy of classifier:  0.893129770992
accuracy of classifier:  0.862595419847
accuracy of classifier:  0.847328244275
accuracy of classifier:  0.859541984733
accuracy of classifier:  0.874809160305
accuracy of classifier:  0.873282442748
accuracy of classifier:  0.862595419847
[Finished in 2.7s]
```

Figure 22. Supervised machine learning accuracy with 3 classes (below 2.0, between 2.0 to 4.0, above 4.0), running 10 times

Furthermore, we also tried to have three movie classes: average score below 2.0 (bad movie), average score above 4.0 (good movie), and average between 2.0 and 4.0 (average movie). In this way, we obtained an around 88% accuracy in prediction using both K-NN and SVM machine learning algorithms (Figure 22).

Therefore, back to business question two, if we have a new movie and all features defined here, we can roughly predict what this movie's score should be located.