IMDB Movie Review Analysis

IST565-Data Mining
Professor Jonathan Fox

By

Daniel Hanks Jr

Executive Summary

The movie industry is an extremely competitive industry in a variety of ways. Not only are movie makers fighting amongst each other for people's dollars, but people themselves want to make an informed decision on which movie to spend their dollars on. There's no indication that that the cost to create movies or the cost to watch movies is going down any time soon. This project proposes to benefit both parties, with a classification system utilizing data mining algorithms to understand the ratings used for movies.

This project involves classifying user rating data based on movie information. The goal would be to be able to predict user rating based on information found in the database. Information available in this database include movie title, genre, actors/actresses, directors, company, year, etc. Questions that could be explored include things like predicting if the movie's genre is sufficient to predict how well a movie will rate among viewers. We should be able to look at that data and say Action movies tend to rate higher than documentaries, for example. Using a system such as this, movie creators should be able to predict which movies should review higher and thus make more money. A consumer should also be able to use this system and be able tell that a certain movie would review higher and most likely be a better investment for them.

The implementation plan for this project is to use data found at the IMDB movie database, which is an extremely popular website for reviewing movies. This will require downloading the necessary datasets, cleaning up that data and preparing it for mining, and ultimately data analysis. The data analysis will consist of interesting and useful visualizations that will show how certain movie classifications can allow us to predict how well a movie will rate, which will let the consumer and movie creators know if it's worth it in the long run.

As the movie industry continues to grow and becomes increasingly competitive it is important that both movie creators and consumers understand what makes a good movie. A good movie equates to a worthy investment for both the creators and viewers of the movie. This makes projects like this, that utilize the latest data mining techniques to help predict what makes a good movies, a worthy investment for all parties involved.

Research Idea

The project I'm proposing involves classifying user rating data based on movie information, specifically the movie genre in this case. We should be able to look at the data and say action movies tend to rate higher than documentaries, and should then be considered a better investment for both movie producers and the audience that invests their money into hopefully a well-made movie.

Data

The data being used is from the IMDB movie database found at www.IMDB.com. This website allows you to download lists like genres, keywords, movies and ratings. This data can be converted into a CSV file where I can clean it up and make the textual information more suitable for data mining. The fields in the spreadsheet included the following:

RefNo – Id for movie (numeric)
title – Title of movie (text)
year- Year movie was released (date)
length-Duration of movie(numeric)
budget-Cost to make movie (numeric)
rating-Numeric score for movie (1.0-10.0)
votes-Amount of people that submitted a ranking of a movie(numeric)
r1-r10-user rankings of movie(numeric)
mpaa – movie rating (text: PG, R, etc)
Action, Animation, Comedy, Drama, Documentary, Romance, Short-Genres(numeric: 1,0 for yes/no)

Data cleanup consisted mostly of removing some columns from the spreadsheet. RefNo and budget fields were the two fields removed. RefNo was simply unneeded and the budget field was missing too much data to be useful.

Analysis

To analyze and mine the dataset we'll use R, which is a software environment for statistical computing and graphics. The dataset is not structured where we can compare the ratings to the different genres. To accomplish this we use a package found in R called reshape2. This package has a built in function that takes the data in a spreadsheet like format and stacks them into a single column of data. With this out of the way, the first thing I'll look at is where the films tend to rate at. Show below is a histogram of IMDB scores. It looks normally distributed with most films between roughly 6 and 7.5 rating.

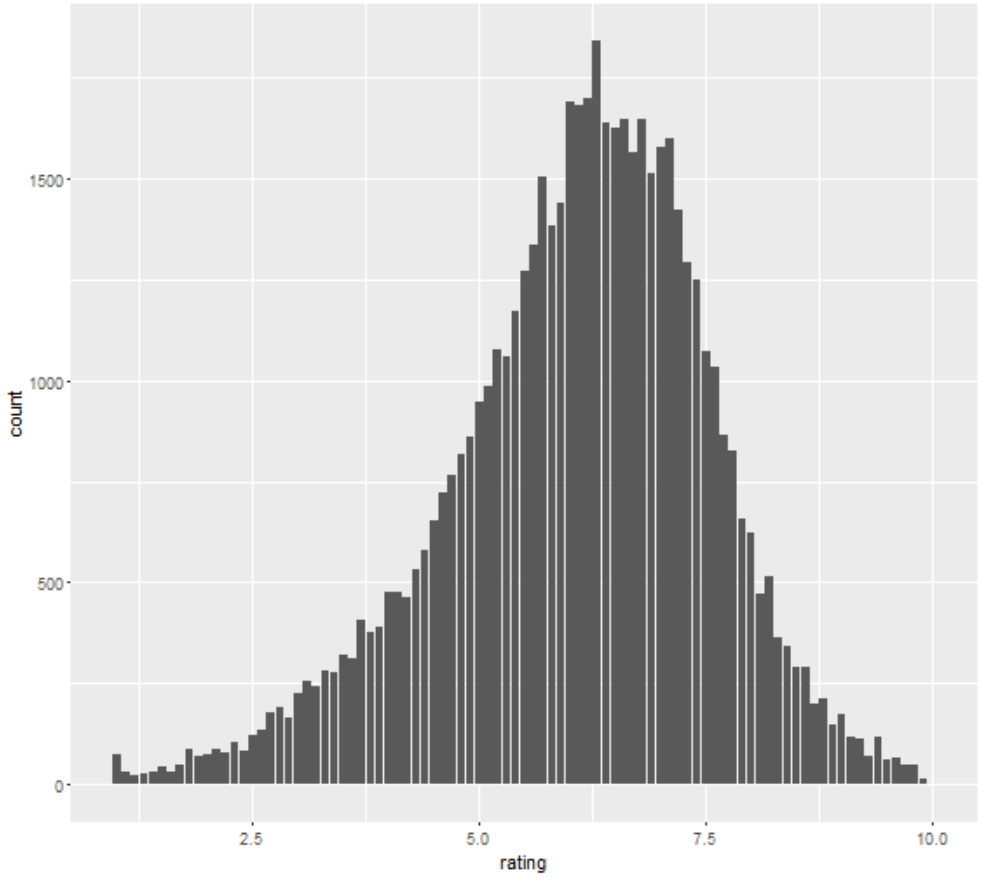

Figure 1

Next, I can look at the distribution of ratings for various genres. This boxplot below shows that documentaries tend to rate the highest, but note that there's a bigger range of discrepancy for a documentary than say an animation rated closely to documentary. What this means is that a documentary or action film tend to have a wider range of opinion which leads to mixed results for scores. An animation film on the other hand, seems to have a very strong possibility of scoring a 6 or above making it a safer choice to invest in.

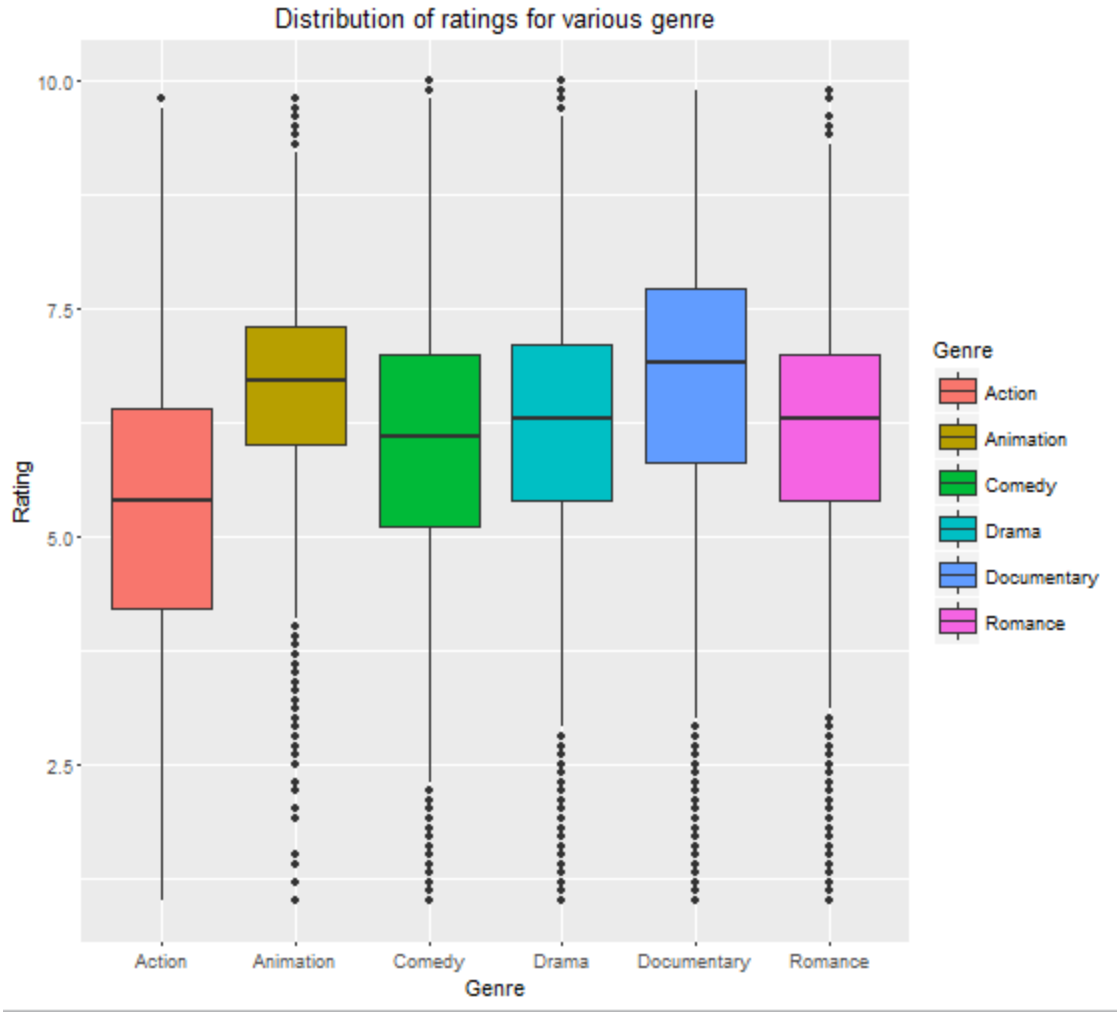

Figure 2

It might also be helpful to note the frequency of movies in each genre to better understand the potential popularity of the genre. The bar graph below shows that while documentaries may tend to rank higher, there's a lot less of them being made. The reasoning for this is most likely due to overall popularity of a documentary being less than other genres. You don't see many documentaries breaking box office records in the summer. The combination of frequency of genre, along with a high ratings among viewers would point towards drama and comedy genres being a safe investment for all parties involved.

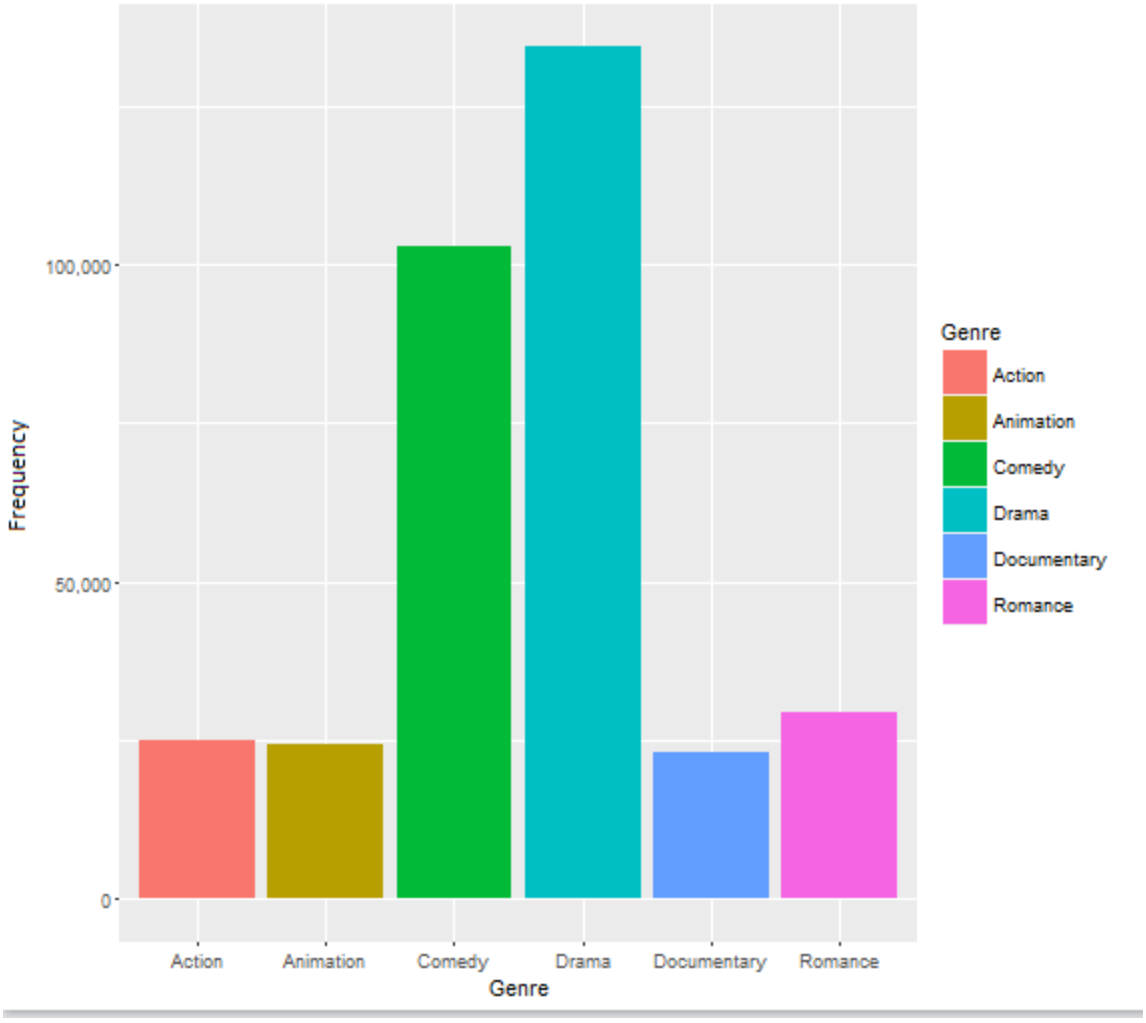

Figure 3

Model

So we now have a feel for what would be a safe genre to pick, but can we create a model that predicts a movies rating based on genre?

Well, with the basic models created so far we know it's pretty tough to predict which genre we should invest in. I can create a linear model to test this further. The model will stick with user ratings and genre. I'll use the movie Independence Day which falls under "Action" for this model. The actual score on the website is a 6.9. My model comes up with the following results:

```
> prediction_ID
       fit      lwr      upr
1 5.292022 5.251432 5.332612
```

About 1.6 points off. Another test of the movie "Interstellar" was over 2 points off. This shows this model needs more variables to possibly become more accurate. I added the 'votes' field in an attempt to tune the algorithm and it didn't change the score.

Recommendation

I believe the data shows that the movie industry is indeed as competitive as anticipated with most movies rating similarly. Analysis of the data shows that the drama and comedy genres tend to review higher and are made more frequently. This makes them safe investments for both viewers and movie creators. Using current data mining techniques we were able to create a model and use it to attempt to make predictions on how movies would score based on their genre. The results of this came in 1.5 – 2 points behind actual movie scores. I could easily use the model with the caveat that the score is +- 1.5 points and predict a score fairly accurately. The problem with this is that the bulk of the scores fall in the 6.0-7.5 range, as show in Figure 1. This indicates just one variable such as genre is not sufficient for predicting the success of a movie as most movies would rank in that +- range. The recommendation is to apply more variables into prediction model which would involve linking even more data such as actors, actresses, directors, studios, etc. Budget would be a significant variable to factor, but as stated before this data simply isn't sufficient to use in the model. Looking into other sources to get more data in this regard would also help in making a more accurate model.

Appendix

Data:

http://imdb.com
ftp://ftp.fu-berlin.de/pub/misc/movies/database
movies.csv-attached

R code used:

```
> head(movies)
                      title year length budget rating votes   r1   r2  r3   r4   r5   r6   r7   r8   r9  r10 mpaa Action
1                       $ 1971    121     NA    6.4   348  4.5  4.5 4.5  4.5 14.5 24.5 24.5 14.5  4.5  4.5            0
2         $1000 a Touchdown 1939     71     NA    6.0    20  0.0 14.5 4.5 24.5 14.5 14.5 14.5  4.5  4.5 14.5           0
3   $21 a Day Once a Month 1941      7     NA    8.2     5  0.0  0.0 0.0  0.0  0.0 24.5  0.0 44.5 24.5 24.5           0
4                 $40,000 1996     70     NA    8.2     6 14.5  0.0 0.0  0.0  0.0  0.0  0.0  0.0 34.5 45.5           0
5 $50,000 Climax Show, The 1975     71     NA    3.4    17 24.5  4.5 0.0 14.5 14.5  4.5  0.0  0.0  0.0 24.5           0
6                   $pent 2000     91     NA    4.3    45  4.5  4.5 4.5 14.5 14.5 14.5  4.5  4.5 14.5 14.5           0
  Animation Comedy Drama Documentary Romance Short
1         0      1     1           0       0     0
2         0      1     0           0       0     0
3         1      0     0           0       0     1
4         0      1     0           0       0     0
5         0      0     0           0       0     0
6         0      0     1           0       0     0
```

#data frame is melted using reshape2 because dataset is not in structure that allows us to compare the distribution of the ratings for various genres

```
> require(reshape2)
Loading required package: reshape2
```

```
movie_data_sub <- movies[, c(1,2,4,5,17,18,19,20,21,22)];
movie_data_sub <- melt(movie_data_sub, c(1,2,3,4));
names(movie_data_sub)[5] <- c("Genre");
movie_data_sub <- subset(movie_data_sub, value == 1);
g_genre <- ggplot(data = movie_data_sub, aes(x = Genre, y = rating, fill = Genre));
g_genre + geom_boxplot() + xlab("Genre") + ylab("Rating") + ggtitle("Distribution of ratings for various genre");
```

```
> movie_data_sub <- movies[, c(1,2,4,5,17,18,19,20,21,22)];
> movie_data_sub <- melt(movie_data_sub, c(1,2,3,4));
> names(movie_data_sub)[5] <- c("Genre");
> movie_data_sub <- subset(movie_data_sub, value == 1);
> g_genre <- ggplot(data = movie_data_sub, aes(x = Genre, y = rating, fill = Genre));
> g_genre + geom_boxplot() + xlab("Genre") + ylab("Rating") + ggtitle("Distribution of ratings for various genre");
```

```
Library(scales)

ggplot(data = movie_data_sub, aes(x = Genre, y = rating, fill=Genre))+geom_bar(stat="identity"
)+scale_y_continuous(labels=comma)

#creation of model
model <- lm(rating ~ Genre, data=movie_data_sub)


#example of model with movie “Interstellar”
movie2<-data.frame(title_type="Interstellar", Genre="Drama", rating=86)
prediction_Interstellar <- predict(model, newdata=movie2, interval="confidence")

prediction_Interstellar
```

```
> prediction_ID
       fit      lwr      upr
1 5.292022 5.251432 5.332612
```

About 1.6 points off. Another test of the movie Interstellar was over 2 points off. This shows this model needs more variables to possibly become more accurate.

```
#summary of model
```

```
> summary(model)

Call:
lm(formula = rating ~ Genre, data = movie_data_sub)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6506 -0.8537  0.1445  0.9463  4.5080

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.29202    0.02071  255.54   <2e-16 ***
GenreAnimation   1.29166    0.03120   41.39   <2e-16 ***
GenreComedy      0.66347    0.02335   28.41   <2e-16 ***
GenreDrama       0.86166    0.02283   37.75   <2e-16 ***
GenreDocumentary 1.35855    0.03175   42.79   <2e-16 ***
GenreRomance     0.87197    0.02920   29.86   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.418 on 55670 degrees of freedom
Multiple R-squared:  0.046,     Adjusted R-squared:  0.04592
F-statistic: 536.9 on 5 and 55670 DF,  p-value: < 2.2e-16
```