## Overview

I was given the task of analyzing the Movie Lens 100k (ml-100k) dataset to produce interesting insight about the data. I created 5 different ranking algorithms and analyzed the results of each and I created a Random Forest Classifier to predict a user’s rating of a given film from an out of sample population.

## Analysis

### Comparison of Ranking Algorithms

I compared 5 different ranking algorithms – 3 to sort of the best movie, and 2 to sort of the most polarizing.

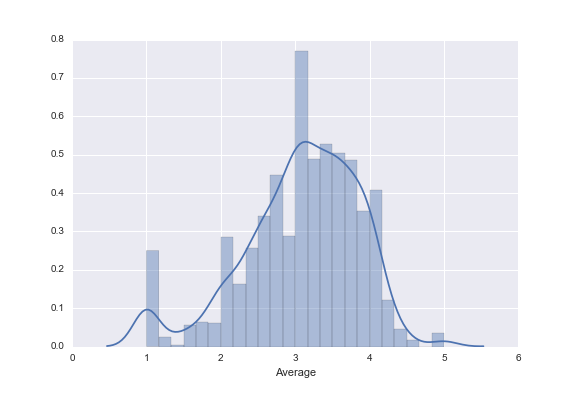
### Best

#### Pure Average

This is not really an algorithm, however, I thought it would be a good starting place just to get an idea of what the data looks like. This method of ranking is clearly flawed in that it does not take into account the number of ratings a film has received, so a film with one five star review will be ranked ahead of a film with 99 five star reviews and one one star review.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Rank | Title | Genre | Year | Average | Ratings |
| 1 | They Made Me a Criminal | Drama | 1939 | 5 | 1 |
| 2 | The Saint of Fort Washington | Drama | 1993 | 5 | 2 |
| 3 | Santa with Muscles | War | 1996 | 5 | 2 |
| 4 | Marlene Dietrich: Shadow and Light | Comedy | 1996 | 5 | 1 |
| 5 | Aiqing wansui | Comedy | 1994 | 5 | 1 |
| 6 | Someone Else's America | Comedy | 1995 | 5 | 1 |
| 7 | Entertaining Angels: The Dorothy Day Story | Comedy | 1996 | 5 | 1 |
| 8 | A Great Day in Harlem | Romance | 1994 | 5 | 1 |
| 9 | Prefontaine | Romance | 1997 | 5 | 3 |
| 10 | Star Kid | Documentary | 1997 | 5 | 3 |

The Distribution of average ratings follows close to a normal distribution with noticeable peaks at the round values (1.0, 2.0, etc.).

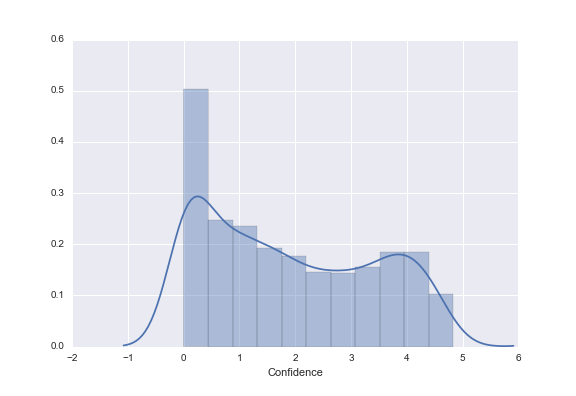


#### Wilson Score

A good ranking algorithm should weight the ratings based on the number of reviews. When thinking about this problem the website Reddit came to mind. Reddit uses what is called the Wilson Score confidence interval for its “Best” ranking. I converted the ratings from each review to a binary value (fresh if the rating > 3, rotten if the rating < 3) and applied the Wilson Score to the aggregated ratings.

This ranking is clearly better than the pure average ranking. The top 10 films, which can be seen below, all have high average rankings and they have all been ranked over 100 times. Since this ranking relies on just the rotten/fresh indicator, the ranking between the very best films is likely inaccurate since the distinction between ratings of 4 and 5 are lost.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Rank | Title | Genre | Year | Average | Wilson Score | Ratings |
| 1 | Casablanca | Noir Mystery | 1942 | 4.46 | 4.84 | 243 |
| 2 | North by Northwest | Comedy Drama | 1959 | 4.28 | 4.82 | 179 |
| 3 | Rear Window | Comedy | 1954 | 4.39 | 4.81 | 209 |
| 4 | The Maltese Falcon | Romance | 1941 | 4.21 | 4.76 | 138 |
| 5 | 12 Angry Men | Sci-Fi | 1957 | 4.34 | 4.75 | 125 |
| 6 | To Kill a Mockingbird | Comedy | 1962 | 4.29 | 4.74 | 219 |
| 7 | The Shawshank Redemption | Comedy Drama | 1994 | 4.45 | 4.73 | 283 |
| 8 | The African Queen | Drama | 1951 | 4.18 | 4.72 | 152 |
| 9 | Vertigo | Thriller | 1958 | 4.25 | 4.70 | 179 |
| 10 | The Manchurian Candidate | Drama | 1962 | 4.26 | 4.69 | 131 |

The rating distribution for this Wilson Score does not display a common pattern. It is heavily skewed towards zero. This is due to the fact that many of the movies have very few reviews which gives them a low Wilson Score.

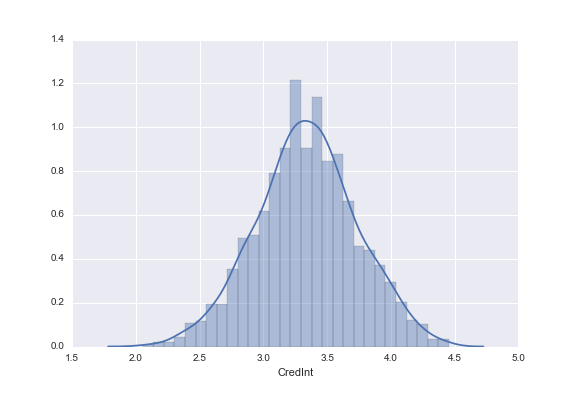
#### Bayesian Credibility Interval

Instead of using the Wilson Score, which assumes that we have no knowledge about films with few reviews and it automatically gives a low score, the information provided in the dataset can be used as a prior to seed the ratings of films with few reviews. I applied an algorithm called the Bayesian Credibility Interval:

Where v is the number of reviews, R is the average rating for the film, C is the average rating for all films within the given film’s genre, and m is a threshold value (I used a value of 10).

This ranking does a better job at accurately ranking the top rated films, because the value approached the actual average as the number of reviews increases.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Rank** | **Title** | **Genre** | **Year** | **Average** | **Ratings** | **BCI** |
| 1 | Casablanca | Noir Mystery | 1942 | 4.46 | 243 | 4.45 |
| 2 | Schindler's List | Romance | 1993 | 4.47 | 298 | 4.44 |
| 3 | A Close Shave | Comedy Drama | 1995 | 4.49 | 112 | 4.42 |
| 4 | The Shawshank Redemption | Comedy Drama | 1994 | 4.45 | 283 | 4.42 |
| 5 | The Wrong Trousers | Comedy Drama | 1993 | 4.47 | 118 | 4.41 |
| 6 | The Usual Suspects | Comedy | 1995 | 4.39 | 267 | 4.35 |
| 7 | Star Wars | Western | 1977 | 4.36 | 583 | 4.35 |
| 8 | Rear Window | Comedy | 1954 | 4.39 | 209 | 4.35 |
| 9 | 12 Angry Men | Sci-Fi | 1957 | 4.34 | 125 | 4.32 |
| 10 | Wallace & Gromit: The Best of Aardman Animation | Documentary | 1996 | 4.45 | 67 | 4.30 |

The distribution of the BCI nearly models a normal distribution. This result occurs because a given film starts with a rating equal to the average of its genre and approaches the raw average for the film as the number of reviews increases.

### Polarizing

Polarizing films an interesting topic to analyze. I created two algorithms to rank films from most polarizing to least.

#### Weighted Up/Down

This ranking is another one inspired by Reddit. As such, I used the rotten/fresh indicator again to create this ranking. The Polar Val is equal to the number of reviews divided by the difference of the absolute value of fresh and rotten reviews (0.75 was used in cases where this value equaled zero).

The most polarizing film received an average rating of exactly 3 despite being reviewed by nearly 150 people.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Rank** | **Title** | **Genre** | **Year** | **Average** | **Ratings** | **Polar Val** |
| 1 | Grumpier Old Men | Comedy | 1995 | 3.04 | 148 | 148.00 |
| 2 | 101 Dalmatians | Thriller | 1996 | 2.91 | 109 | 109.00 |
| 3 | Hoodlum | Thriller | 1997 | 2.93 | 73 | 97.33 |
| 4 | The Nutty Professor | Comedy | 1996 | 2.91 | 163 | 81.50 |
| 5 | The Man Who Knew Too Little | Action Horror Sci-Fi | 1997 | 2.92 | 52 | 69.33 |
| 6 | The Preacher's Wife | War | 1996 | 2.93 | 68 | 68.00 |
| 7 | Evita | Comedy | 1996 | 2.98 | 259 | 64.75 |
| 8 | Star Trek: The Motion Picture | Action Adventure | 1979 | 3.03 | 117 | 58.50 |
| 9 | Nine Months | Comedy Drama | 1995 | 2.93 | 58 | 58.00 |
| 10 | Thanks for Everything! Julie Newmar To Wong Foo | Comedy | 1995 | 2.89 | 57 | 57.00 |

#### Weight by Rating

This ranking was made in an attempt to utilize the full set of 1-5 ratings for each film. I divided the number of 1, 2, 4, and 5 ratings by the weighted difference of 1 and reviews, 2 and 4 reviews, and 3 reviews. This was done in a way such that a film with more 1 and 5 reviews would be ranked ahead of a film with more 2 and 4 reviews and a film with nearly all 3 reviews would be at the end of the list.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Rank** | **Title** | **Genre** | **Year** | **Average** | **Ratings** | **Polar Val** |
| 1 | Year of the Horse | Horror | 1997 | 3.29 | 7 | 7.00 |
| 2 | C'est arrivé près de chez vous | Action Adventure | 1992 | 3.00 | 4 | 5.33 |
| 3 | The Newton Boys | Drama | 1998 | 3.00 | 4 | 5.33 |
| 4 | Head Above Water | Children | 1996 | 3.00 | 4 | 5.33 |
| 5 | Hugo Pool | War | 1997 | 3.40 | 5 | 5.00 |
| 6 | Best Men | Crime | 1997 | 3.40 | 5 | 5.00 |
| 7 | The Old Lady Who Walked in the Sea | Horror | 1991 | 2.60 | 5 | 5.00 |
| 8 | Mina Tannenbaum | Drama | 1994 | 3.67 | 6 | 3.00 |
| 9 | The Stars Fell on Henrietta | Comedy Drama | 1995 | 2.33 | 3 | 3.00 |
| 10 | Captives | Drama | 1994 | 2.33 | 3 | 3.00 |

This is a link to the full dataset of all the rankings for each algorithm.



## Random Forest Classifier to Predict Film Rankings

I created a Random Forest Classifier model using Scikit Learn to predict the rating that a user would give a movie based on a set of features which describe the user and the film.

#### Data Treatment

The only descriptive info provided about the films was the title, release date, genre. In an effort to improve the model’s accuracy of prediction I retrieved the runtime data for the films from the Online Movie Database (OMDB). The information about the genre of each film was provided as flags in columns corresponding to each genre. I converted these 19 columns into a single Genre column which held a string representation of the genre (i.e. “Comedy”, “Action Adventure”, “Romance”, etc.). I converted the categorical variables to integer labels using scikit learn’s preprocessing library and the Label Encoder function.

Training

I trained the model using the u1.base datasets provided in the ml-100k data. I joined this data, which contained just the rating, to the user and item data which I head preprocessed. The list of features I used for this model were: Genre, Runtime, Year, Age, Gender, and Occupation.

#### Result

I tested the model using the u1.test data which resulted in an average error of 1.1807.

The exact error and number of instance for each were recorded in the table below.

|  |  |
| --- | --- |
| **Predicted Value Missed By:** | **n** |
| 0 | 5287 |
| ± 1 | 8409 |
| ± 2 | 4238 |
| ± 3 | 1627 |
| ± 4 | 439 |

#### Conclusion

The accuracy of this model can be greatly improved. Given the time, I would have sought out more sources of external data to join with the film information, such as budget, total profits, language, and possibly other sources of ratings. I would also have liked to use geographic data gathered from the zip code provided in the user dataset to see if geographic location of the user would improve the predictive capabilities of the model.