DATA 607: Week 11

Cindy Lin

2025-04-09

INTRODUCTION

Week 11 assignment's goal is to learn about recommender's system. The most common example of this is the recommendations that Netflix or amazon provides based on the choices selected prior. For this assignment we will be using the Global Baseline Estimate which outputs the predicted rating. The formula is adding the global average rating and the user bias and the item bias. The bias will be deviation of the rating.

Loading library

I am loading the tidyverse library because there are functions that can help tidy the loaded data

Loading the data

```
get_data <- read.csv("movierating.csv", header = TRUE)

get_data
```

```
##    PersonID                Movies Rating
## 1         1          Inside Out 2     NA
## 2         2          Inside Out 2    5.0
## 3         3          Inside Out 2     NA
## 4         4          Inside Out 2     NA
## 5         5          Inside Out 2     NA
## 6         1  Deadpool & Wolverine    3.0
## 7         2  Deadpool & Wolverine    5.0
## 8         3  Deadpool & Wolverine    3.0
## 9         4  Deadpool & Wolverine    4.0
## 10        5  Deadpool & Wolverine     NA
## 11        1                Wicked     NA
## 12        2                Wicked     NA
## 13        3                Wicked     NA
## 14        4                Wicked     NA
## 15        5                Wicked     NA
## 16        1           Beetlejuice    1.0
## 17        2           Beetlejuice    2.0
## 18        3           Beetlejuice    1.0
## 19        4           Beetlejuice     NA
## 20        5           Beetlejuice     NA
```

```
## 21        1 Venom: The Last Dance     NA
## 22        2 Venom: The Last Dance    2.0
## 23        3 Venom: The Last Dance    2.0
## 24        4 Venom: The Last Dance     NA
## 25        5 Venom: The Last Dance     NA
## 26        1        Dune: Part Two     NA
## 27        2        Dune: Part Two     NA
## 28        3        Dune: Part Two     NA
## 29        4        Dune: Part Two     NA
## 30        5        Dune: Part Two     NA
## 31       NA                       2.8
```

Load the csv file from my working directory and viewing the data.

Global Average Rating

```
global_avg_rating <- mean(get_data$Rating, na.rm = TRUE)

print(global_avg_rating)
```

```
## [1] 2.8
```

2.8 is the global average rating, which means that the total average of all users and movies.

User Bias

```
user_bias <- get_data %>%
  filter(!is.na(Rating)) %>%
  group_by(PersonID) %>%
  summarize(b_u = mean(Rating - global_avg_rating), .groups = "drop")

user_bias
```

```
## # A tibble: 5 x 2
##   PersonID   b_u
##      <int> <dbl>
## 1        1  -0.8
## 2        2   0.7
## 3        3  -0.8
## 4        4   1.2
## 5       NA   0
```

The user bias is the users rating relative to the global average so how much it deviates. This is done by adding the average of each users and minus the global average. For example, id 1 had an user average of 2 (which means their average of what they rated in all the movies), with the global average = 2.8 (2 - 2.8 = -0.8).

Movie Bias

```
movie_bias <- get_data %>%
  filter(!is.na(Rating)) %>%
  group_by(Movies) %>%
  summarize(b_i = mean(Rating - global_avg_rating), .groups = "drop")

movie_bias
```

```
## # A tibble: 5 x 2
##   Movies                    b_i
##   <chr>                   <dbl>
## 1 ""                       0
## 2 "Beetlejuice"           -1.47
## 3 "Deadpool & Wolverine"   0.95
## 4 "Inside Out 2"           2.2
## 5 "Venom: The Last Dance" -0.8
```

The movie bias is the movie rating relative to the global average so how much it deviates. This is done by adding the average of each movie and minus the global average. For example, Beetlejuice had an user average of 1.333, with the global average = 2.8 (1.333 - 2.8 = -1.47).

Predicted Value

```
predicted_value <- get_data %>%
  left_join(user_bias, by = "PersonID") %>%
  left_join(movie_bias, by = "Movies")

predicted_value <- predicted_value %>%
  mutate(
    b_u = ifelse(is.na(b_u), 0, b_u),
    b_i = ifelse(is.na(b_i), 0, b_i),
    predicted_rating = 2.8 + b_u + b_i
  )

predicted_value
```

```
##    PersonID                Movies Rating  b_u       b_i predicted_rating
## 1         1          Inside Out 2     NA -0.8  2.200000        4.2000000
## 2         2          Inside Out 2    5.0  0.7  2.200000        5.7000000
## 3         3          Inside Out 2     NA -0.8  2.200000        4.2000000
## 4         4          Inside Out 2     NA  1.2  2.200000        6.2000000
## 5         5          Inside Out 2     NA  0.0  2.200000        5.0000000
## 6         1  Deadpool & Wolverine    3.0 -0.8  0.950000        2.9500000
## 7         2  Deadpool & Wolverine    5.0  0.7  0.950000        4.4500000
## 8         3  Deadpool & Wolverine    3.0 -0.8  0.950000        2.9500000
## 9         4  Deadpool & Wolverine    4.0  1.2  0.950000        4.9500000
## 10        5  Deadpool & Wolverine     NA  0.0  0.950000        3.7500000
## 11        1                Wicked     NA -0.8  0.000000        2.0000000
## 12        2                Wicked     NA  0.7  0.000000        3.5000000
```

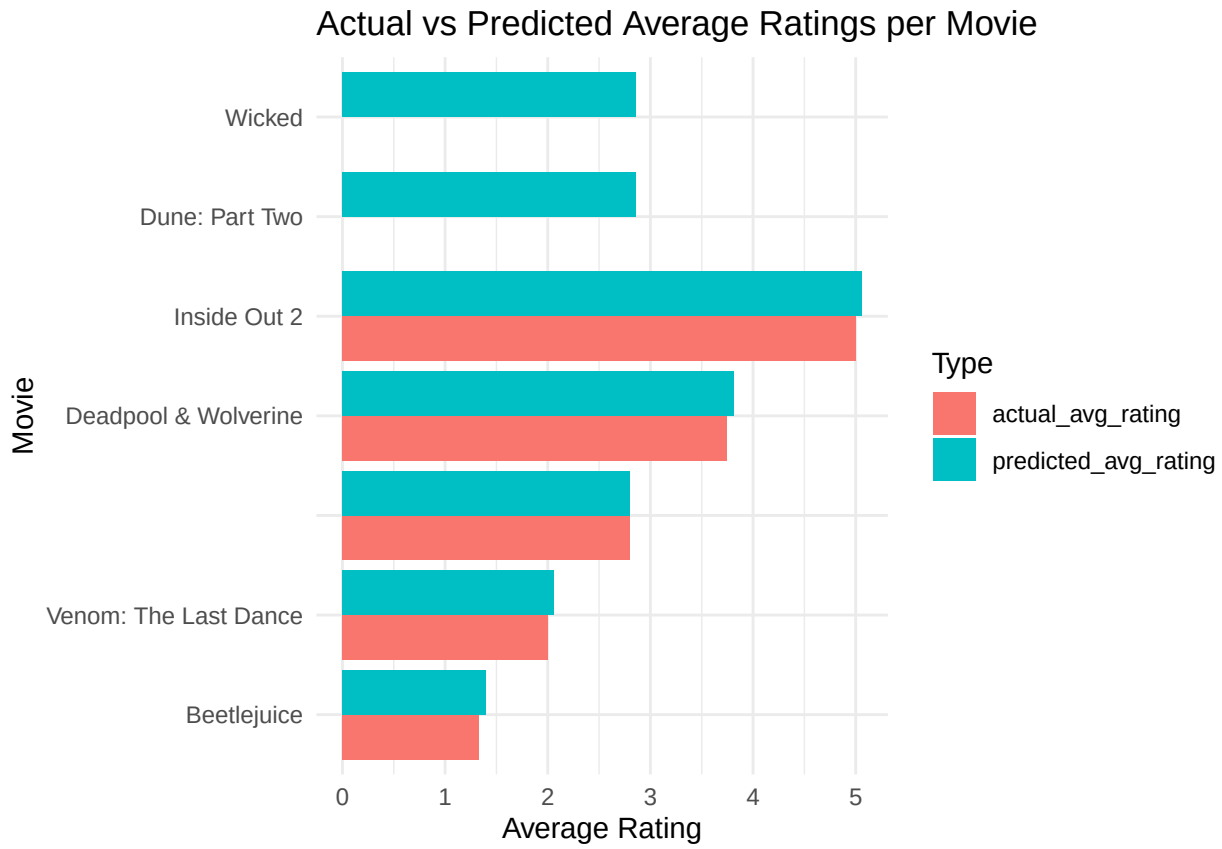
```
## 13        3                Wicked     NA -0.8  0.000000        2.0000000
## 14        4                Wicked     NA  1.2  0.000000        4.0000000
## 15        5                Wicked     NA  0.0  0.000000        2.8000000
## 16        1           Beetlejuice    1.0 -0.8 -1.466667        0.5333333
## 17        2           Beetlejuice    2.0  0.7 -1.466667        2.0333333
## 18        3           Beetlejuice    1.0 -0.8 -1.466667        0.5333333
## 19        4           Beetlejuice     NA  1.2 -1.466667        2.5333333
## 20        5           Beetlejuice     NA  0.0 -1.466667        1.3333333
## 21        1 Venom: The Last Dance     NA -0.8 -0.800000        1.2000000
## 22        2 Venom: The Last Dance    2.0  0.7 -0.800000        2.7000000
## 23        3 Venom: The Last Dance    2.0 -0.8 -0.800000        1.2000000
## 24        4 Venom: The Last Dance     NA  1.2 -0.800000        3.2000000
## 25        5 Venom: The Last Dance     NA  0.0 -0.800000        2.0000000
## 26        1        Dune: Part Two     NA -0.8  0.000000        2.0000000
## 27        2        Dune: Part Two     NA  0.7  0.000000        3.5000000
## 28        3        Dune: Part Two     NA -0.8  0.000000        2.0000000
## 29        4        Dune: Part Two     NA  1.2  0.000000        4.0000000
## 30        5        Dune: Part Two     NA  0.0  0.000000        2.8000000
## 31       NA                         2.8  0.0  0.000000        2.8000000
```

Using the global baseline estimate formula, predicted value was added. For those with ratings already, we can see the difference of how far and close the actual value is and for those that do not have the values, we can then use the predicted value. The prediction comes from what we know about the user and about the movie as well as the average of everyone else.

Actual and Predicted

```
rating_comparison <- predicted_value %>%
  group_by(Movies) %>%
  summarize(
    actual_avg_rating = mean(Rating, na.rm = TRUE),
    predicted_avg_rating = mean(predicted_rating, na.rm = TRUE),
    n = sum(!is.na(Rating))
  )

rating_comparison_long <- rating_comparison %>%
  select(Movies, actual_avg_rating, predicted_avg_rating) %>%
  pivot_longer(cols = c(actual_avg_rating, predicted_avg_rating),
               names_to = "Type", values_to = "Rating")

rating_comparison_clean <- rating_comparison_long %>%
  filter(!is.na(Movies))


ggplot(rating_comparison_long, aes(x = reorder(Movies, Rating), y = Rating, fill = Type)) +
  geom_col(position = "dodge") +
  coord_flip() +
  labs(title = "Actual vs Predicted Average Ratings per Movie",
       x = "Movie", y = "Average Rating") +

  theme_minimal()
```

```
## Warning: Removed 2 rows containing missing values or values outside the scale range
## (`geom_col()`).
```



I wanted to see the actual and predicted comparison so I used ggplot. For the most part, the predicted and actual are pretty close with the predicted being more generous but not by much. Wicked and Dune: Part Two are empty so only the user bias is used for those predicted rating, unlike the others where users and movie bias is used.

Conclusion

This was an interesting exercise as I can see it being applicable in a lot of cases. While we probably encountered this in the business world on products, I can see it also being applicable in other under-served area. With the formula being focus on two parts (items and users), I think an interesting project could be aligning this to course recommendations in higher education. While majors are supposed to be prescribed, there are usually a lot of room for flexibility, especially if we are talking about undergraduate degrees. I think having some recommender model with courses (with a similar model as amazon or Netflix) would be interesting to see.