data607_assignment3A

Joao

2025-09-09

Overview

This report intends to apply the Global Baseline Estimate algorithm to the collected data after surveyed 5 people on 6 movie ratings. The goal is to preditc a person's rating based on their other ratings and the movie average ratings. Despite being a general prediction technique is still a reasonable approach for predictions.

```
ratings <- read_csv("https://raw.githubusercontent.com/JDO-MSDS/DATA-607-Week3/refs/heads/main/Assignmen
```

Reshape and clean

```
# cleaned by fix quotes in all columns
names(ratings) <- names(ratings) %>%
  str_replace_all("[\u2018\u2019]", "") %>%
  str_squish()

# I used print(names(ratings)) to check if 1st column is 'Critic' and check the quotes issue

#reshaping table with long to have each critic's data per row

long <- ratings %>%
  pivot_longer(-Critic, names_to = "movie", values_to = "rating")

# only rows wth ratings
obs <- long %>%
  filter(!is.na(rating))
```

Calculate mean, user and movie deviation

```
# movie ratings mean using obs since it has no NA
movies_mean <- obs %>%
  summarize(movies_mean = mean(rating)) %>%
  pull(movies_mean)

# user deviation/bias
critic_deviation <- obs %>%
  group_by(Critic) %>%
  summarize(critic_mean = mean(rating), .groups = "drop") %>%
```

```r
  mutate(c_d = critic_mean - movies_mean) %>%
  select(Critic, c_d)

# each movie deviation from the mean
movie_deviation <- obs %>%
  group_by(movie) %>%
  summarize(movie_mean = mean(rating), .groups = "drop") %>%
  mutate(m_d = movie_mean - movies_mean) %>%
  select(movie, m_d)
```

Missing rates prediction

```r
predict_user <- function(user_name, top_n = 3) {
  # select user row if user exists
  user_row <- ratings %>% filter(Critic == user_name)
  if (nrow(user_row) == 0) stop("User can't be found")

  # unrated movies by the user
  missing_movies <- user_row %>%
    pivot_longer(-Critic, names_to = "movie", values_to = "rating") %>%
    filter(is.na(rating)) %>%
    select(movie)

  if (nrow(missing_movies) == 0) {
    return(tibble(movie = character(), pred = numeric()))
  }

  # user bias/dev
  cd <- critic_deviation %>%
    filter(Critic == user_name) %>%
    pull(c_d)

  # predictions
  preds <- missing_movies %>%
    left_join(movie_deviation, by = "movie") %>%
    mutate(m_d = replace_na(m_d, 0),
           pred = movies_mean + cd + m_d) %>%
    arrange(desc(pred)) %>%
    slice_head(n = top_n)

  preds
}
```

```r
predict_user("Finn", 5)
```

```
## # A tibble: 2 x 3
##   movie            m_d  pred
##   <chr>          <dbl> <dbl>
## 1 Lilo & Stitch -0.208  3.79
## 2 Heretic       -0.458  3.54
```

```r
predict_user("Garrett", 5)
```

```
## # A tibble: 1 x 3
##   movie         m_d  pred
##   <chr>       <dbl> <dbl>
## 1 Spider-Man 0.0417  4.24
```

```r
predict_user("Hayley", 5)
```

```
## # A tibble: 0 x 2
## # i 2 variables: movie <chr>, pred <dbl>
```

```r
predict_user("Stephen", 5)
```

```
## # A tibble: 2 x 3
##   movie          m_d  pred
##   <chr>        <dbl> <dbl>
## 1 Friendship   0.292  4.04
## 2 Inside Out 2 0.292  4.04
```

```r
predict_user("Walter", 5)
```

```
## # A tibble: 1 x 3
##   movie            m_d  pred
##   <chr>          <dbl> <dbl>
## 1 Lilo & Stitch -0.208  4.19
```

```r
critics <- ratings %>% dplyr::pull(Critic) %>% unique()

all_preds <- purrr::map_df(
  critics,
  ~ predict_user(.x, top_n = Inf) %>% dplyr::mutate(Critic = .x, .before = 1)
)

pred_table <- all_preds %>%
  dplyr::mutate(pred = round(pred, 2)) %>%
  dplyr::arrange(Critic, dplyr::desc(pred)) %>%
  dplyr::select(Critic, movie, pred)

knitr::kable(pred_table, align = "l",
             caption = "Predicted for unrated movies")
```

Table 1: Predicted for unrated movies

Critic	movie	pred
Finn	Lilo & Stitch	3.79
Finn	Heretic	3.54
Garrett	Spider-Man	4.24
Stephen	Friendship	4.04

Critic	movie	pred
Stephen	Inside Out 2	4.04
Walter	Lilo & Stitch	4.19

Conclusion

The predictions table shows all the predictions for unrated movies by all users. Despite not being an individualized and advanced prediciton tecnhique, the GBE ended up getting pretty solid predictions.