Movie Capstone

2024-09-13

R Markdown

Data Loading

We start by loading the necessary libraries and importing the MovieLens dataset.

```
# Load necessary libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
# Load the data
ratings <- read.csv("ratings.csv")
movies <- read.csv("movies.csv")
```

Data Preperation

The ratings and movies datasets are merged, and the data is split into training (edx) and validation (final_holdout_test) sets.

```r
# Merge the ratings and movies datasets on the movieId column
merged_data <- merge(ratings, movies, by = "movieId")

# Split the data into edx (90%) and final_holdout_test (10%)
set.seed(1)
test_index <- createDataPartition(merged_data$rating, p = 0.1, list = FALSE)
edx <- merged_data[-test_index, ]
final_holdout_test <- merged_data[test_index, ]

# Ensure final_holdout_test has only users and movies that are also in edx
final_holdout_test <- final_holdout_test %>%
  semi_join(edx, by = "movieId") %>%
  semi_join(edx, by = "userId")
```

Model and Evaluation

We calculate the average rating for each movie as the baseline model and then calculate RMSE to evaluate its performance.

```r
# Calculate the average rating for each movie
movie_avg_rating <- edx %>% group_by(movieId) %>% summarise(avg_rating = mean(rating))

# Merge the average ratings into the final_holdout_test set
final_holdout_test <- final_holdout_test %>%
  left_join(movie_avg_rating, by = "movieId")

# Calculate RMSE between actual ratings and predicted average ratings
rmse <- RMSE(final_holdout_test$avg_rating, final_holdout_test$rating)
print(paste("Baseline RMSE:", rmse))
```

```
## [1] "Baseline RMSE: 0.96170576181978"
```

Conclusion

The baseline model produced an RMSE of r round(rmse, 4). Future work will involve more advanced models to improve the predictions.

Explanation:

- **Introduction Section**: Explains the goal of the project.
- **Data Loading Section**: Loads the necessary libraries and dataset.
- **Data Preparation Section**: Merges the datasets, splits the data, and ensures consistency between `edx` and `final_holdout_test`.
- **Model and Evaluation Section**: Calculates the average rating, merges it, and computes the RMSE.
- **Conclusion Section**: Displays the RMSE and mentions possible future steps.