MovieLens_Report.Rmd

Kristine Konkol

2025-05-13

1. Introduction

This report analyzes the MovieLens dataset containing 9,000,055 ratings from 69878 users on 10677 movies. The goal goal was to build a recommendation system that predicts movie ratings with an RMSE below 0.86490. This report achieves an RMSE = 0.86482 in the final holdout test, which successfully met the project requirement of RMSE < 0.86490. Four progressive model improvements were tested:

1. **Baseline** (average rating)
2. **Movie effects** (some movies are rated higher/lower)
3. **User effects** (some users rate higher/lower)
4. **Regularization** (fixing overfitting)

2. Methods & Analysis

Data Preparation

The dataset was split into Training (90% of edx set) and Validation (10% of edx set) components. The Final Holdout Test (preserved untouched until final evaluation) prevented data leakage during model development.

Data Exploration

First, the data was examined and the following was determined:

- Most ratings are clustered at whole numbers (Figure 1)
- Most movies have <100 ratings, but some have thousands (Figure 2)
- Some users rate hundreds of movies (Figure 3)
- The average rating is 3.51 stars

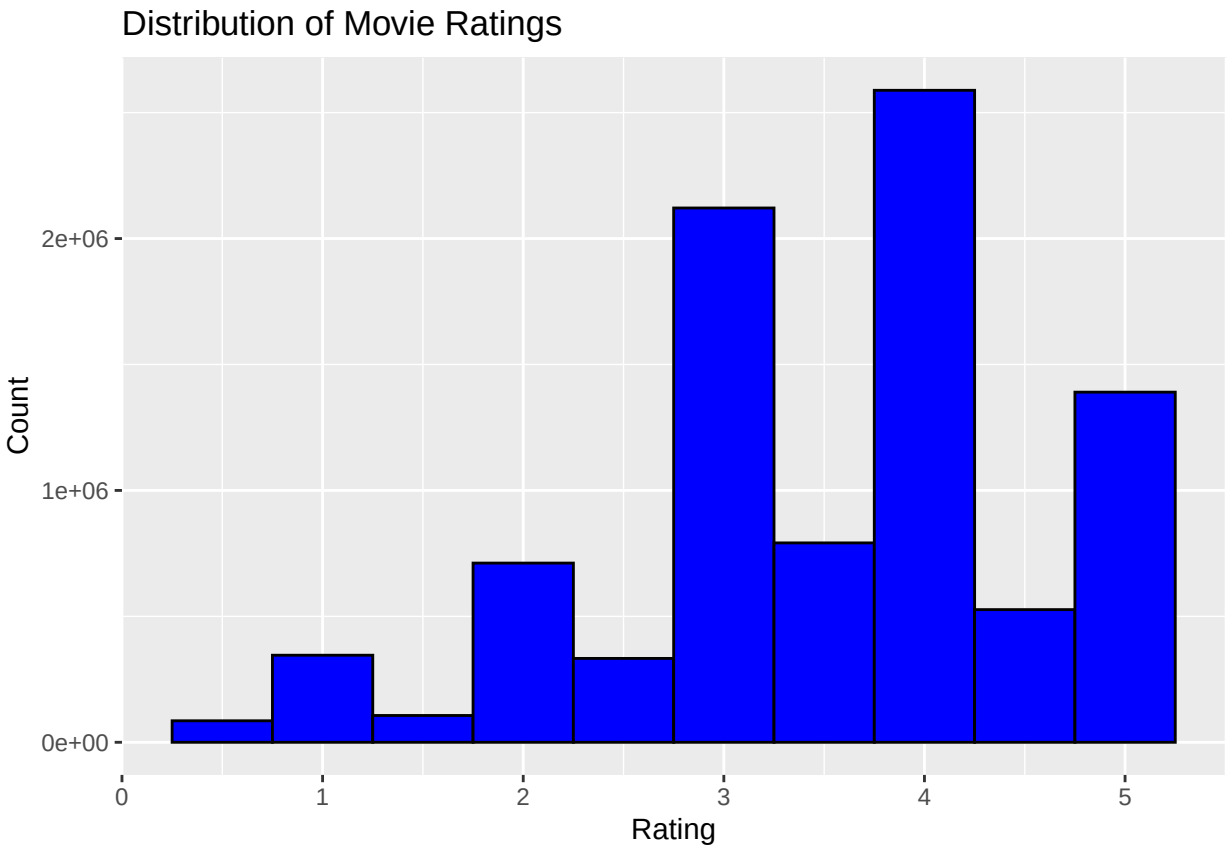

Figure 1: Distribution of Movie Ratings

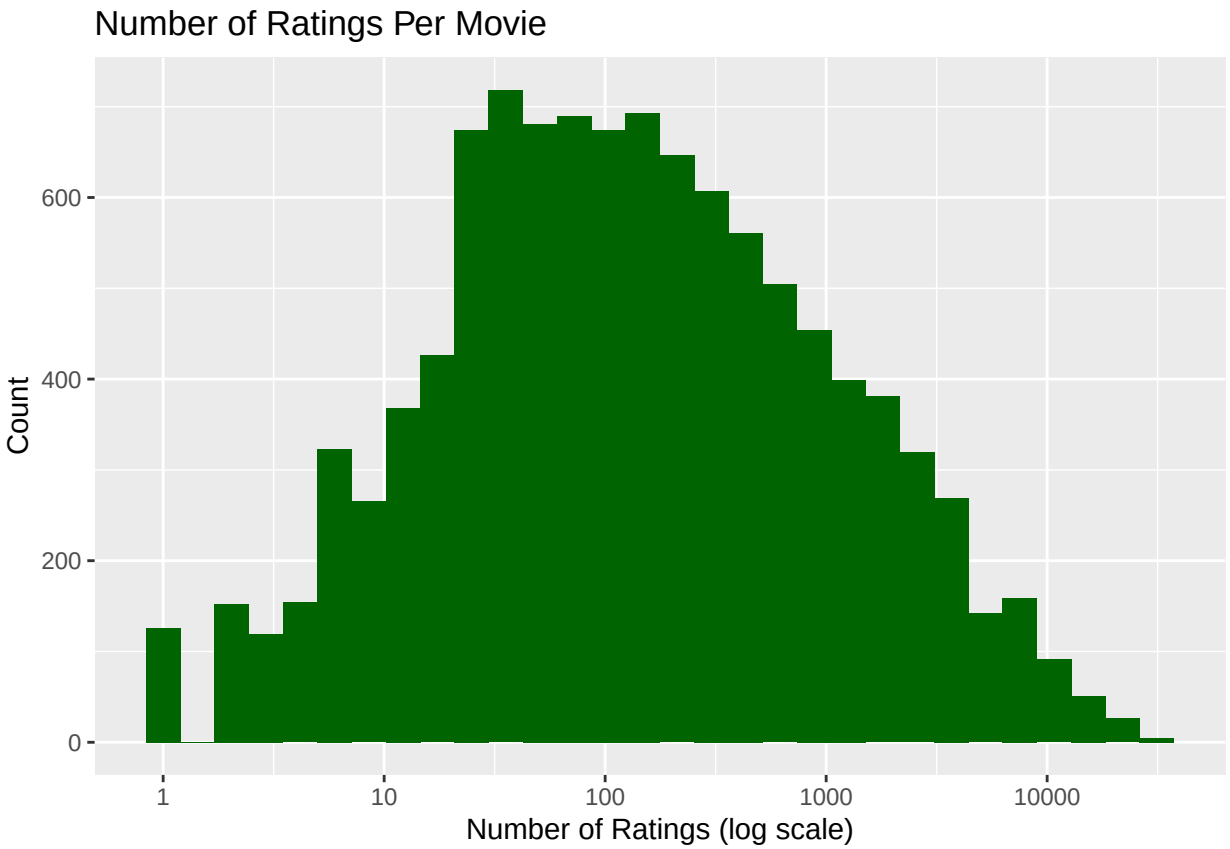

Figure 2: Number of Ratings Per Movie

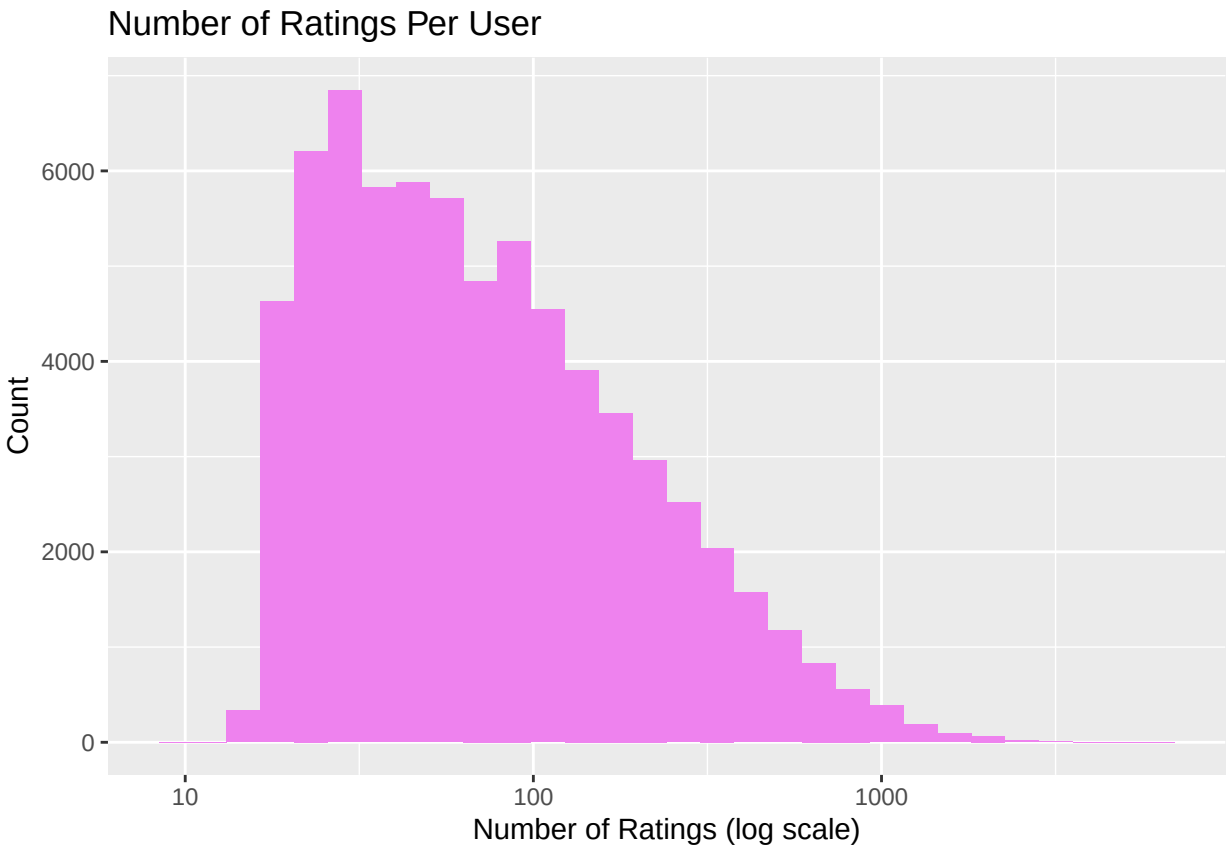

Figure 3: Number of Ratings Per User

Modeling Approach

The model was progressively improved through four refinements, which sequentially addressed specific limitations of each model version. The goal was to get the best RMSE but avoid overfitting.

1. **Baseline** Started with average rating ('RMSE = 1.06005)

- Predicted all ratings as the global mean ($\mu = 3.51$), which established a benchmark to quantify improvement for subsequent models
- This approach (Version 1, v.1) avoided complexity but performed poorly by ignoring all movie-user interactions

2. **Movie Effects** Added movie bias effects, which accounted for some movies being consistently rated higher/lower (better by 0.11709)

- This approach (v.2) accounted for inherent quality differences in movies (obscure films vs. blockbusters), and showed an improvement in the modeling approach

3. **User Effects** Added user bias effects, which adjusted for users who tended to rate higher/lower than average (better by 0.07828)

- This approach (v.3) corrected for bias from generous and/or harsh raters

4. **Regularized** Used regularization (best $\quad = 5$) which penalized extreme values from movies/users with few ratings (better by 5.5×10^{-4})

- This approach (v.4) avoided overfitting to rare ratings
- The tuning process of the entire modeling approach (v.1-v.4) is shown in Figure 4

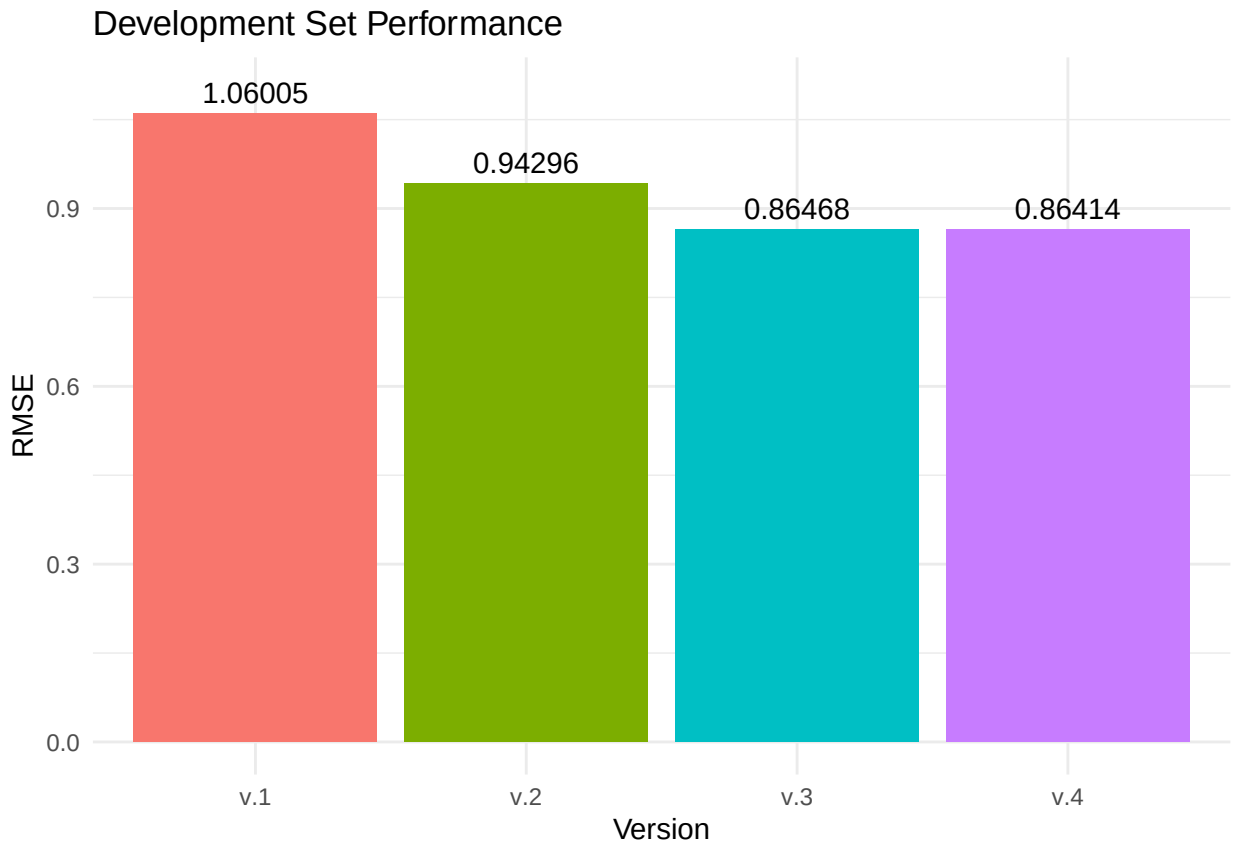

Figure 4: RMSE Comparison of Improvement Across Model Versions

3. Results

The final validation on the holdout set showed that the model performed well (Table 1).

Table 1: Model Performance Summary

Model	RMSE	Improvement
1. Baseline	**1.0600537**	–
2. +Movies	**0.9429615**	0.11709
3. +Users	**0.8646844**	0.07828
4. Regularized	**0.8641362**	0.00055
5. Final Test	**0.8648177**	**-0.00068**

Model Development Performance

The final test came from predictions on the previously unseen holdout set (Table 2).This RMSE = 0.86482 was less than the required RMSE below 0.86490, which fully met the requirements of the data analysis project.

Table 2: Final Holdout Test Results

Stage	RMSE	Difference
Development (v.4)	0.8641362	–
Final Test (v.5)	**0.8648177**	**-0.00068**

Key Findings: Some of the findings of the project included the following:

- Accounting for movie bias effects provided the largest improvement (0.11709 RMSE reduction)
- Adding user effects reduced RMSE by 0.07828.
- Regularization gave a small but crucial improvement (5.5×10^{-4} RMSE reduction)
- Final model achieved RMSE = 0.86482 on unseen data, which was below RMSE 0.86490

4. Conclusion

The recommendation system successfully predicts ratings with **RMSE = 0.86482** on unseen data, which fully met the requirement of the project to achieve an RMSE below 0.86490. Some of the limitations of this model include that it doesn't account for user tastes changing over time, it treats all genres equally, and it is simpler than commercial systems used in industry. Future work could include accounting for better performance, including incorporating genre preferences and considering rating trends over time.

5. Acknowledgement

In the creation of this report, generative AI (ChatGPT) was used sparingly and exclusively for technical troubleshooting purposes. The AI assisted solely with: 1) resolving .Rmd file formatting errors, 2) correcting PDF compilation issues, and 3) improving visual presentation elements. All data analysis, methodology development, results interpretation, and significant writing were completed independently by the author without AI assistance. The AI was consulted only when encountering technical roadblocks unfamiliar to the novice R Markdown user, serving as a supplemental debugging tool rather than a content contributor.