analysis-report

Ganesh Shelke

15/06/2019

Introduction

Recommendation systems use user behavior to predict or suggest recommendations of other items to the user. For example Facebook Amazon, Google and Netflix use these algorithms to customize recommendations and thus increase revenue.Recently there are many companies working in many fields and use these systems to tackle the problem. For example, a healthcare software company can use the patients' data to predict if the patient has a cancer and if so, is it benign or malignant. This helps in early detection of cancer and eventually save many lives.

Objective

In this project, we will showcase a movie recommendation algorithm to predict user ratings of movies based on a database of user reviews.The goal is to develop a model for movie rating given 5 base variables (user ID, movie title, movie year, movie genre, and review date). We have used 10M MovieLens dataset. And final algorithm was used to predict ratings on a validation set. The performance of the algorithm was evaluated by RMSE, or root mean square error in stars between predicted rating and actual rating.

The Dataset

We have used the 10M version of the MovieLens dataset. The MovieLens data set contains 10000054 rows, 10677 movies, 797 genres and 69878 users.The edx dataset (train set) contains 9,000,055 ratings of 10,677 movies by 69,878 users and consisted of 90% of the original benchmark MovieLens 10M dataset while validation set contains 999,999 ratings.

The dataset consists of the following variables:

"userId", "movieId", "rating", "timestamp", "title", "genres"

- `userId`: Unique user ID number.
- `movieId`: Unique movielens movie ID number.
- `rating`: User-provided ratings on a 5-star scale with half-star increments starting from 0.5
- `timestamp`: Time of user-submitted review in epoch time,
- `title`: Movie titles including year of release as identified in IMDB
- `genres`: A pipe-separated list of film genres

Performance

Exploratory data analysis on the modeling data revealed correlations between rating and movie ID, user ID and genre. The final model accounting for movie, user and genre bias yielded a root mean squared error (RMSE) of 0.865 on the validation set.

Methods

Data Download and Extraction of Validation Set

We downloaded the MovieLens 10M dataset from https://grouplens.org/datasets/movielens/10m/ and then processed by code provided by the course where we removed a 10% validation set from the initial data. The validation set was used only to evaluate the final model. This study uses the remaining 90% of the data, edx data, for all training and testing.

Exploratory Data Analysis (EDA)

We performed exploration was on the edx dataset. Selected results from the exploratory analysis are reported for variables used in the final model: movie, user and base genre.

Data Wrangling

We have performed some wrangling steps on both the edX and validation sets before splitting it into training and test data and the following variables were generated (with its type in bracket)

- `action` - whether film includes "Action" genre (logical)
- `adventure` - whether film includes "Adventure" genre (logical)
- `animation` - whether film includes "Animation" genre (logical)
- `children` - whether film includes "Children" genre (logical)
- `comedy` - whether film includes "Comedy" genre (logical)
- `crime` - whether film includes "Crime" genre (logical)
- `documentary` - whether film includes "Documentary" genre (logical)
- `drama` - whether film includes "Drama" genre (logical)
- `fantasy` - whether film includes "Fantasy" genre (logical)
- `filmNoir` - whether film includes "Film-Noir" genre (logical)
- `horror` - whether film includes "Horror" genre (logical)
- `imax` - whether film includes "IMAX" genre (logical)
- `musical` - whether film includes "Musical" genre (logical)
- `mystery` - whether film includes "Mystery" genre (logical)
- `romance` - whether film includes "Romance" genre (logical)
- `sciFi` - whether film includes "Sci-Fi" genre (logical)
- `thriller` - whether film includes "Thriller" genre (logical)
- `war` - whether film includes "War" genre (logical)
- `western` - whether film includes "Western" genre (logical)
- `unknown` - whether film includes "Unknown" genre (logical)
- `ratingFactor` - star rating as a factor (factor)

Training and test sets

The EdX 10M MovieLens data were split into a 90% training set (edx dataset) and 10% test set (validation data). The test set contained only movies and users also present in the training set.

Modeling

We trained a linear model on a 90% training set and tested on a 10% test set. The covariates were iteratively added and evaluated with root mean squared error (RMSE) as the loss function. The final model accounted for movie, user and genre effects, regularized by number of reviews per movie, user or genre. The final model yielded an RMSE of 0.862 on the test set and 0.865 on the validation set.

Results

Exploratory Data Analysis

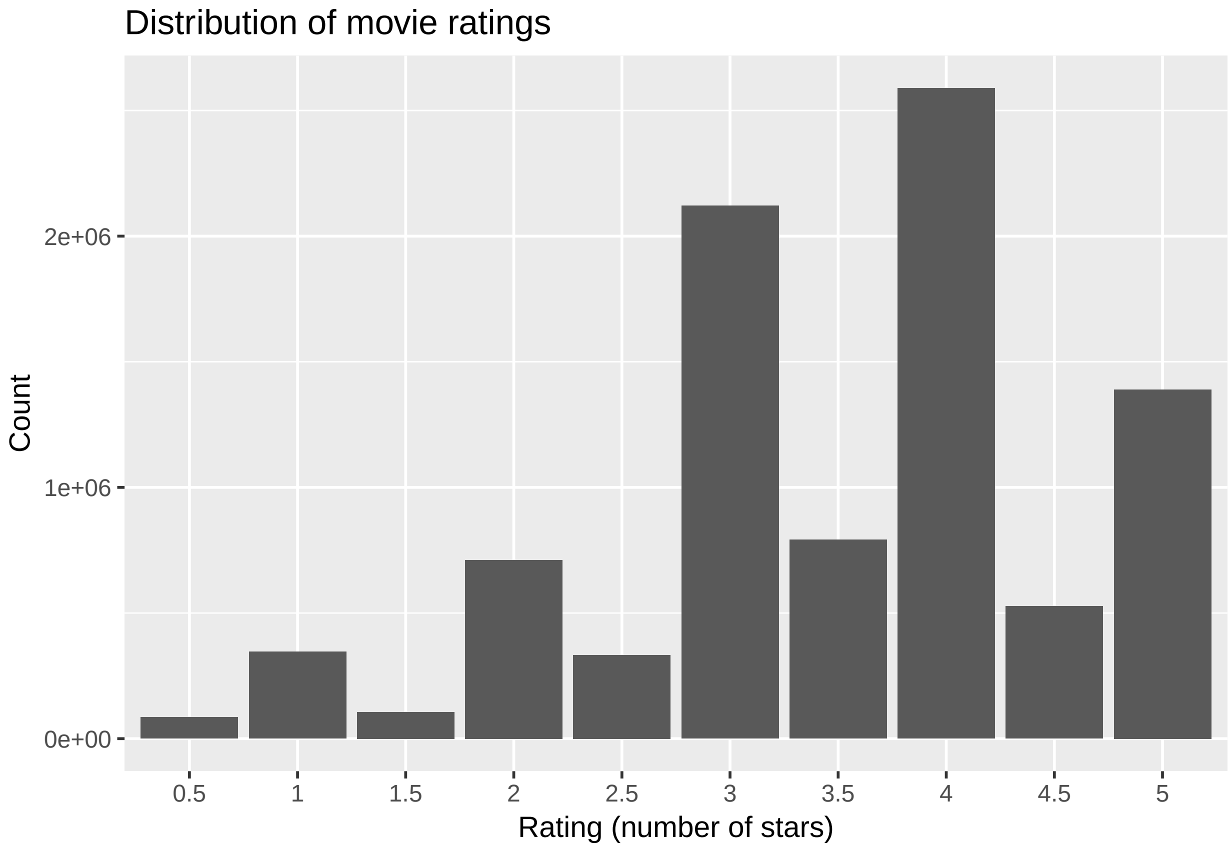
The properties of covariates relevant to the final model are described here.

```
## Warning in set.seed(1, sample.kind = "Rounding"): non-uniform 'Rounding'
## sampler used
```

Rating: The `rating` variable

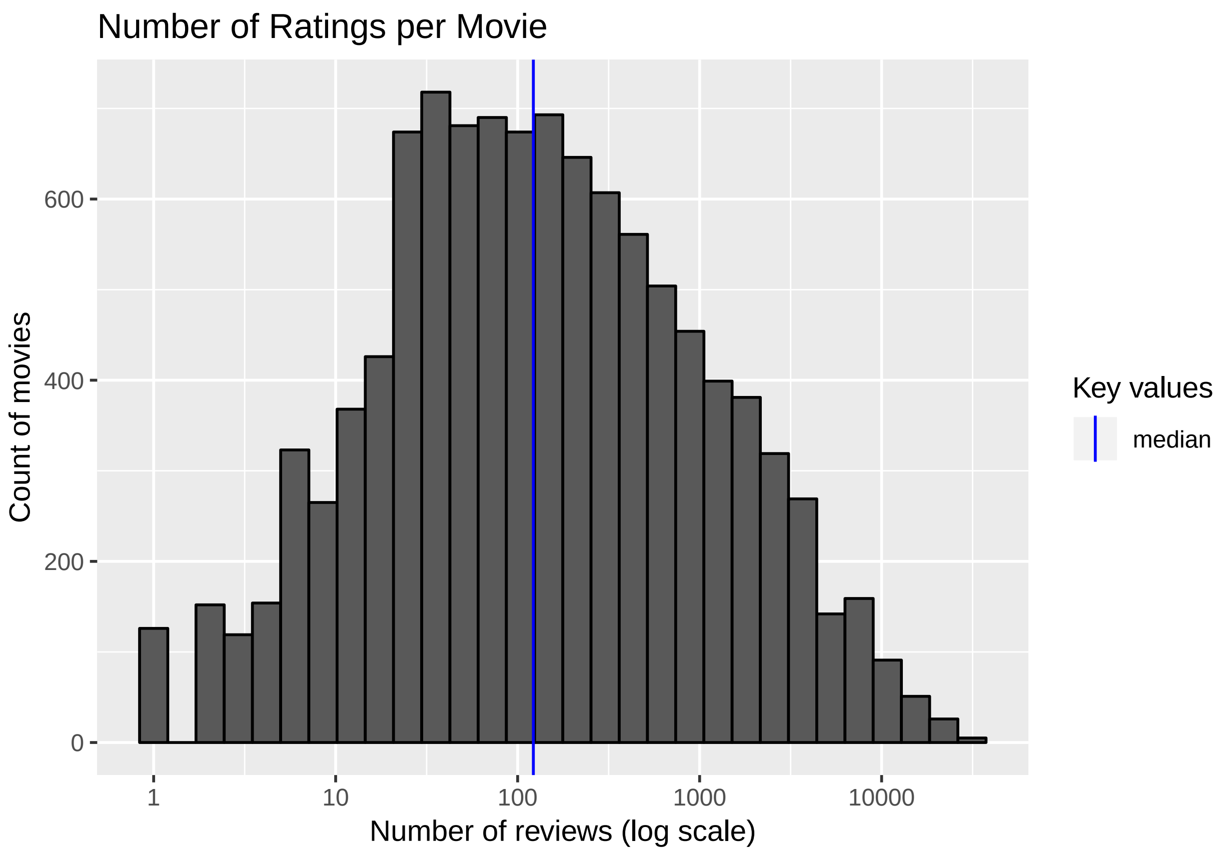
The rating variable consists of a numeric five-star rating with half-star increments. The average rating across all movies and users is 3.51 stars. There is a clear discretization effect where whole-star ratings are more frequent than half-star ratings. The distribution is skewed to the right with a mode of 4 stars.

```
## [1] 3.512465
```



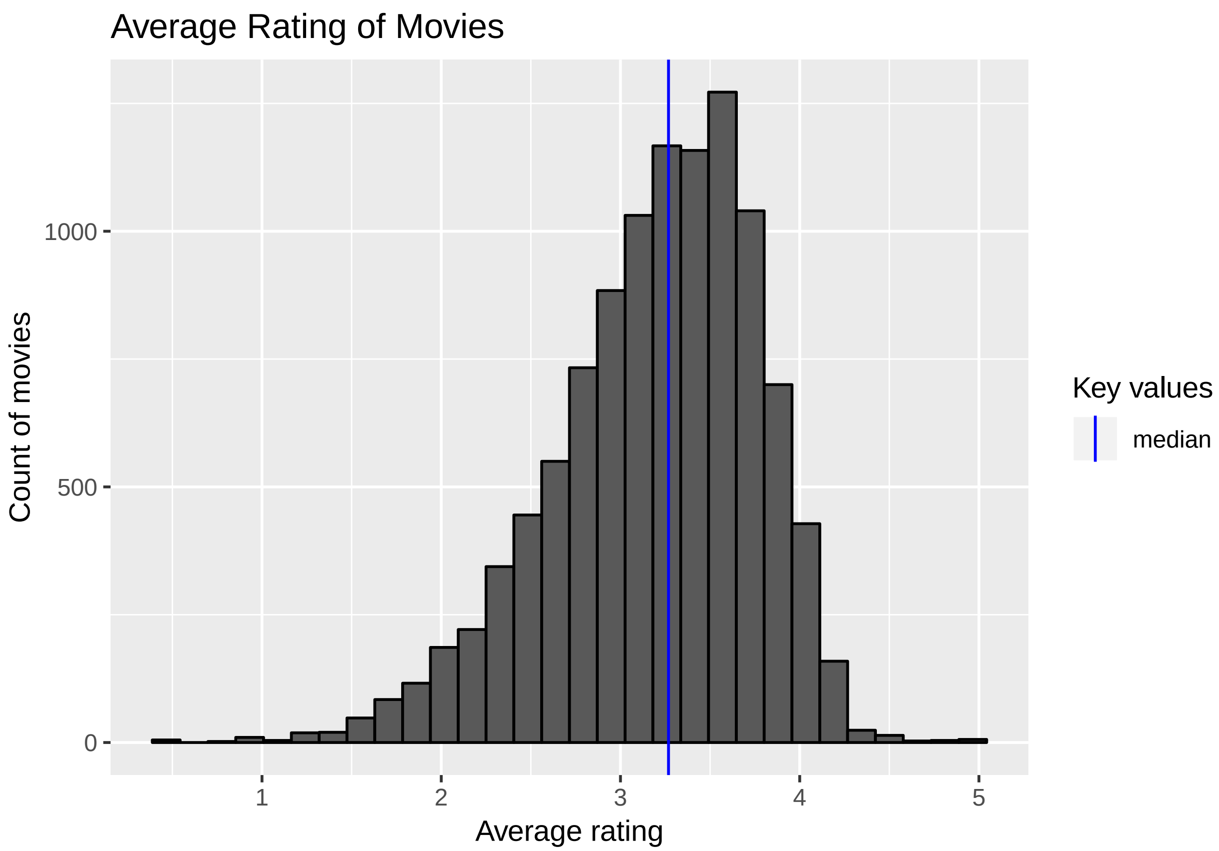
Movies: The `movieId` Variable

The movie variable `movieId` contains the unique MovieLens movie ID number for each movie. A given `movieId` value is always paired with the same `title` value. We have 10677 unique movie IDs in the dataset with a median of 122 ratings.



Average Rating Per Movie

The average mean rating per movie is 3.192, much lower than the average rating over all movies of 3.512 and the average rating over users of 3.614. This discrepancy suggests that movies with higher number of reviews get the bulk of positive reviews.



Relationship Between Movie Rating (stars) and Number of Movie Reviews

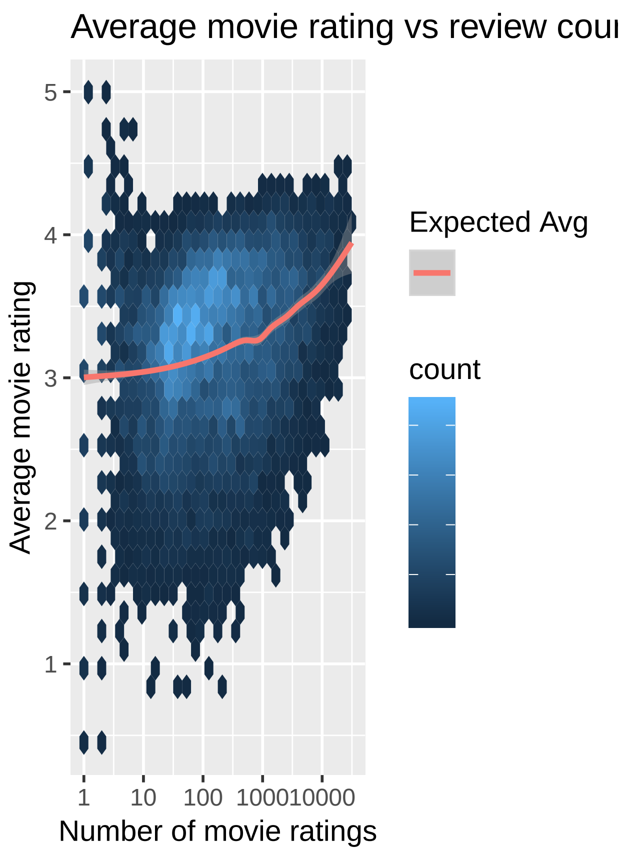
Movies with more ratings tend to have higher average ratings. Also, the number of movies with extremely large or extremely small standard deviations decreases as the number of reviews increases. This suggests that movie ratings stabilize over time and movies with many reviews have more trustworthy expected ratings than movies with few reviews.

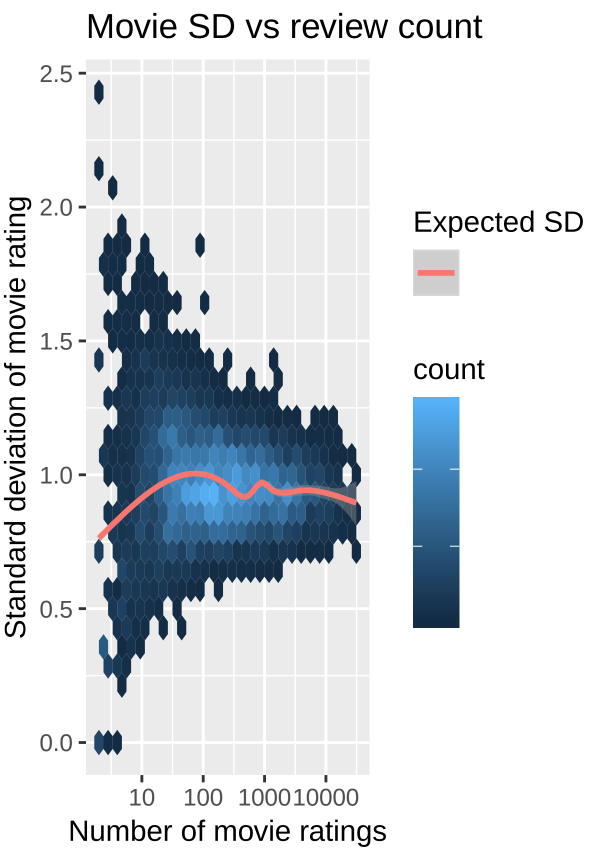
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 126 rows containing non-finite values (stat_binhex).
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```

```
## Warning: Removed 126 rows containing non-finite values (stat_smooth).
```



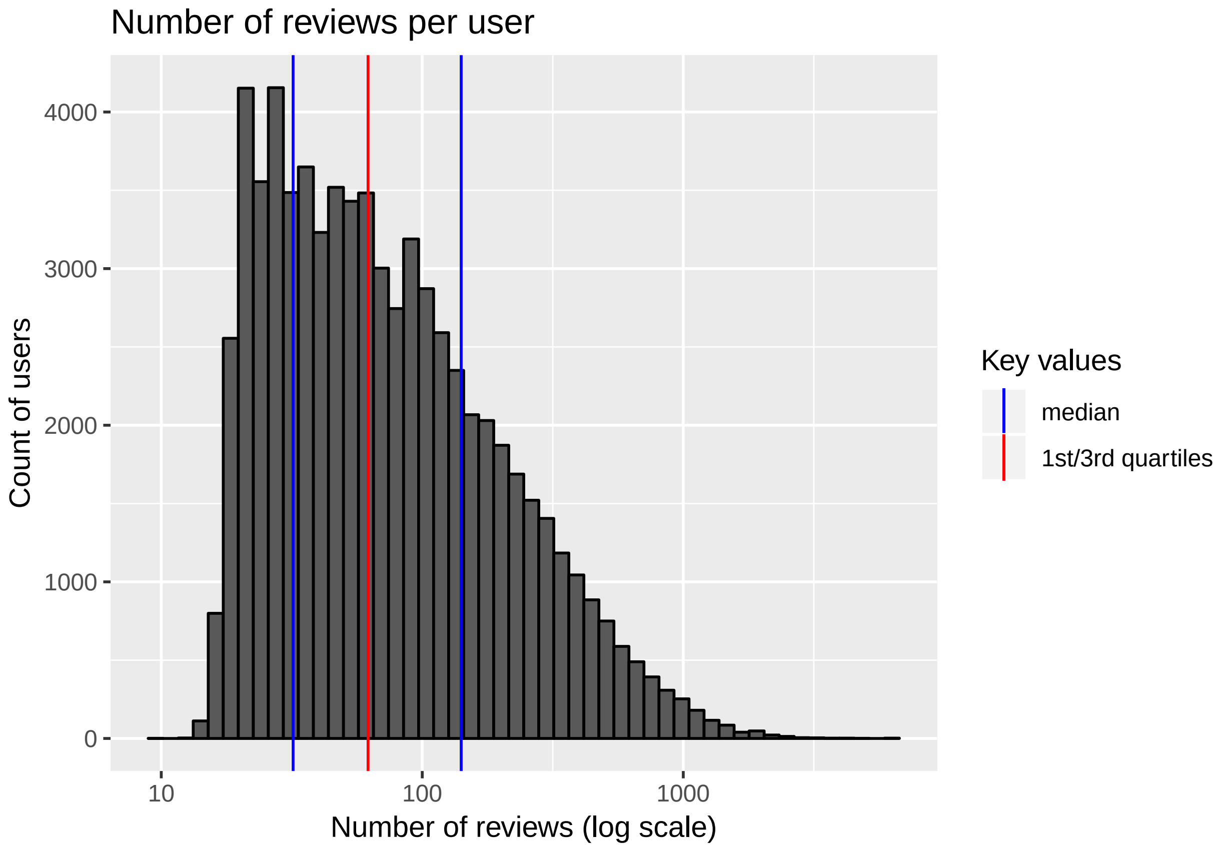


Users: The `userId` variable

The user variable consists of a `userId`, a unique integer ID number for each user that can be converted to a factor. There are 69878 unique users in the dataset.

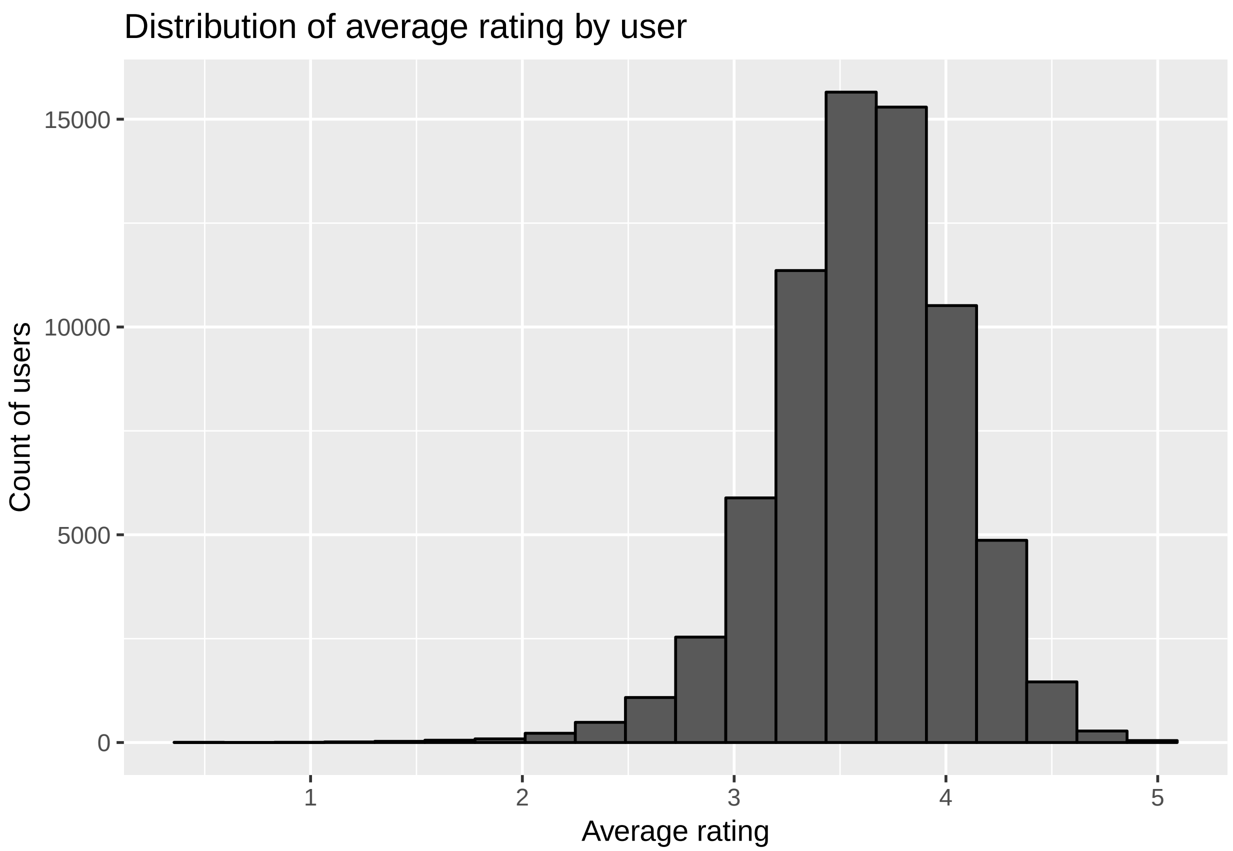
Number of User Reviews

All users have at least 10 reviews with a median of 62 reviews. The interquartile range was 32-141 reviews.



Average User Review

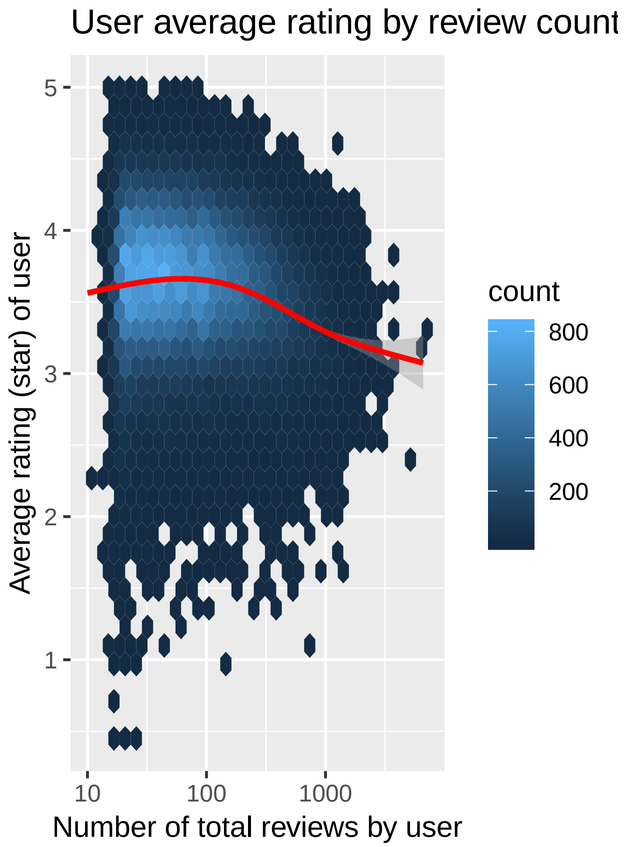
Different users have different average ratings. The mean average rating across all users is 3.614, but individual users have distinct rating distributions and innate tendencies to rate higher or lower.

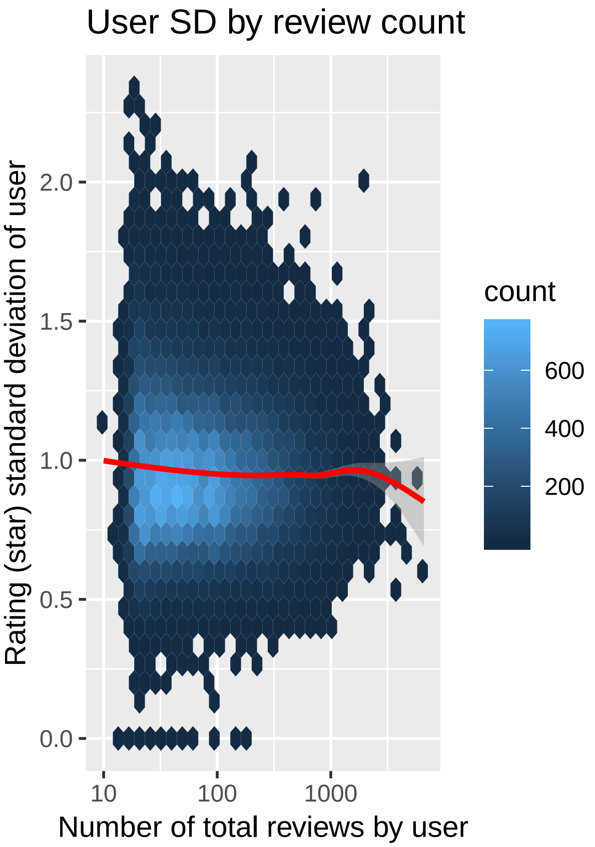


Relationship Between Number of Total Reviews by User and User Ratings

Multiple effects are visible when comparing average user review and standard deviation of user reviews to number of reviews per user. First, the majority of users tend to have a mean review around 3.5, but the expected rating decreases as number of reviews by user increases. Also, users with low numbers of reviews are more likely to have extreme average ratings over 4.5 or below 2. Extremely large or small user standard deviations also tend to become less frequent as the number of reviews increases, suggesting users may become more consistent in rating over time and that user average may be more predictive for users with higher numbers of reviews.

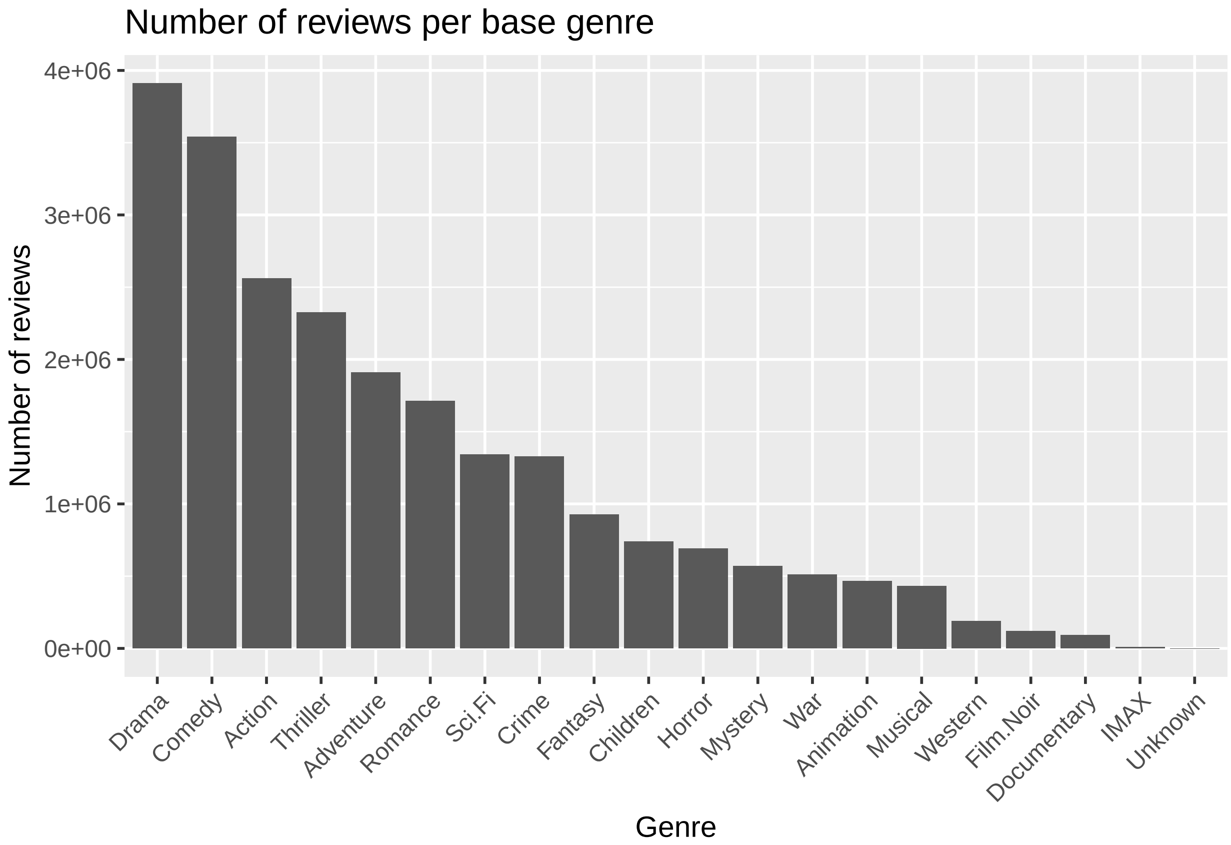
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```





Genres: The `genres` variable

`genres` is a pipe-separated list of film genres that apply to a given movie. There are 19 base genres plus an unknown category. In order to facilitate handling of movies with multiple genres, and to avoid excessive modeling bias by making genre groups too specific and small, the genres are reduced to the 19 base genres plus the unknown genre category. The most popular genres by review count are Drama (3.91 million reviews) and Comedy (3.54 million reviews) and the least popular genres are IMAX (8181 reviews) and Unknown (7 reviews).



Average ratings of base genres

Genres differ greatly in their average rating and the standard deviation of their ratings. Film-Noir has the highest average rating with the lowest standard deviation, while Horror has the lowest average rating with the highest standard deviation.

```
## Joining, by = "genre"
```

```
## Warning: Column `genre` joining character vector and factor, coercing into
## character vector
```

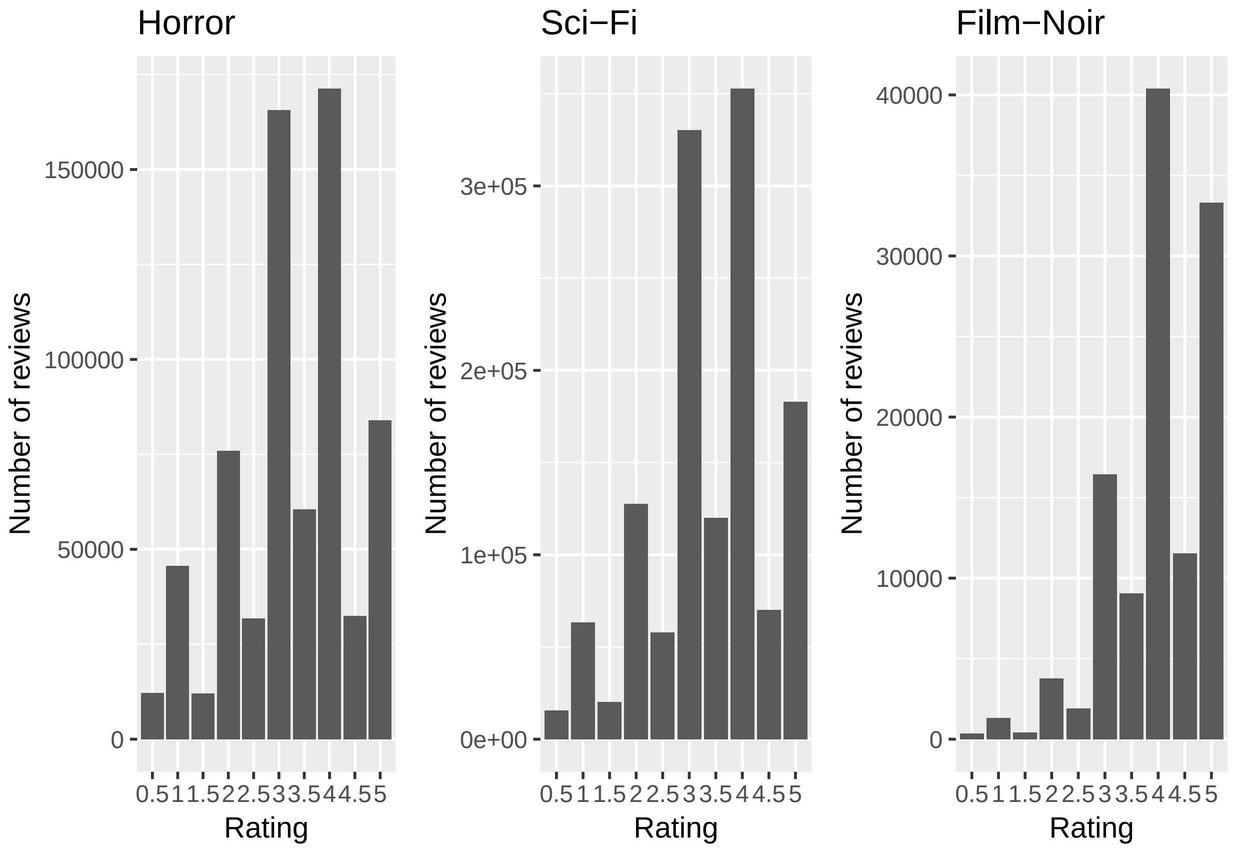
Table 1: Rating trends of different base genres

Genre	Reviews	Average Rating	Rating SD
Film-Noir	118541	4.011625	0.8871659
Documentary	93066	3.783487	1.0038488
War	511147	3.780813	1.0118036
IMAX	8181	3.767693	1.0323171
Mystery	568332	3.677001	1.0002628
Drama	3910127	3.673131	0.9953970
Crime	1327715	3.665925	1.0119106
Unknown	7	3.642857	1.1073349
Animation	467168	3.600644	1.0193206
Musical	433080	3.563305	1.0568704

Genre	Reviews	Average Rating	Rating SD
Western	189394	3.555918	1.0237553
Romance	1712100	3.553813	1.0304136
Thriller	2325899	3.507676	1.0311492
Fantasy	925637	3.501946	1.0654636
Adventure	1908892	3.493544	1.0529353
Comedy	3540930	3.436908	1.0746511
Action	2560545	3.421405	1.0665828
Children	737994	3.418715	1.0923977
Sci-Fi	1341183	3.395743	1.0927764
Horror	691485	3.269815	1.1499549

Distributions of base genres

Different base genres have different rating distributions. Consider the comparison between horror, sci-fi and film-noir, where Horror and Sci-Fi have a relative left skew and film-noir has a relative right skew.



Modeling

For all models, the mean overall rating, movie, user or genre effects were calculated using the training set. The model's performance was evaluated on the test set using RMSE as the loss function. The final model containing movie, user and genre effets was then applied to the training set.

Table 2: Root mean squared error (RMSE) of regularized models on test and validation sets

Model	RMSE
Mean rating	1.0593
Regularized movie effect added	0.9412
Regularized user effect added	0.8623
Regularized genre effect added	0.8620
Final model on validation set	0.8649

The naive assumption - predicting the mean rating

The simplest possible recommendation system is to predict the same rating for all movies and all users. This assumes there is a true rating that applies to all users and movies, and that all variation we see is random error. The naive assumption generates a test set RMSE of 1.06.

Movie effects with regularization

Each movie has an intrinsic quality reflected in the difference between its average rating and the average rating across all movies.Because different movies have different numbers of ratings, and the number of ratings affects confidence in the movie's true quality, regularization was employed for estimates of the movie effect b_i. The regularized movie bias was added to the naive prediction of the mean rating `muHat`. Addition of the movie effect reduced the test set RMSE to 0.941.

User effect with regularization

Each user has a different distribution of ratings. Each user's intrinsic bias towards or away from the mean can be determined by calculating the mean difference between a user's rating and the expected rating for a given movie as determined in our previous model. A further improvement to our model includes this user effect as b_u. The regularized user bias was added to the regularized movie model. Addition of the user effect reduced the test set RMSE to 0.8624.

Genre effects with regularization

Each genre has a different distribution of ratings. Each genre's intrinsic bias towards or away from the mean can be determined by calculating the difference between a genre's mean rating and the expected rating for a given movie as determined in our previous model. A further improvement to our model includes a genre effect $g_{u,i}$. The regularized genre bias was added to the regularized movie-user model. Addition of the genre effect decreased the test set RMSE to 0.8621.

Validation set performance of final model

The final model accounting for regularized movie bias, regularized user bias and regularized genre bias yielded an RMSE of 0.865.

Conclusion

The final model generated predictions on the validation set with an RMSE of 0.865. This model incorporated regularized movie effects, regularized user effects and regularized genre effects. The final RMSE corresponds to an average error of 0.865 stars out of 5, or 17.3%. This suggests the predictions are actionable and useful for a recommendation system, but also that results could be improved with further modeling.