HarvardX_PH125.9x_Capstone_MovieLens_Project

Janalin Black

5/21/2021

1. Introduction

1.1 Assignment

The purpose of this project is to train a machine learning algorithm to predict movie ratings. The expectation is to improve on the data analysis strategies used within this course series related to recommendation systems. Specifically, the goal is to go beyond a recommendation system that includes a regularized model with movie and user effects. This machine learning algorithm uses the typical error loss, residual mean squared error (RMSE), on a test set where success is achieved when the final RMSE is at or below .86490.

1.2 DataSet

The movie recommendation system is created using the MovieLens dataset found in the dslabs package. For ease of computation, the 10M version of the MovieLens dataset is used. Assignment directions included code that split the 10M MovieLens data into two datasets, edx and validation. The edx dataset is used to train and test the movie ratings predictor and the validation dataset is used to evaluate the RMSE of the final algorithm.

Because the validation dataset is only used to evaluate the final predictive model, the edx dataset is divided into separate test and train sets to design and test potential predictive models with the test set including 10% of the edx dataset.

1.3 Key Steps to Predictive Model

The process in creating the final predictive model includes analyzing the train dataset for potential predictors beyond movie and user effects. Modeling is accomplished using the train and test sets and a final RMSE number is produced using the validation data as the final hold-out dataset.

The successful predictive model expands movie and user predictors to include separate genre effects. Additionally, regularization is used on all predictors to improve the overall accuracy and achieve successful RMSE numbers.

2. Methods and Analysis

2.1 Data Cleaning

The initial review of the edx dataset shows 6 columns and over 9 million observations with no missing values. For all separated datasets, the timestamp column needs to be converted to readable form and separate columns created for month, day, hour, and year. The movie title also includes release year and

should be divided into two columns. The genres column includes several classifications (e.g. comedy, drama, sci-fi) per row; it is arguable that each group of genre combinations is a category and therefore, unique. However, it is also plausible that each classification is unique and should be separated for better predictive potential. Because of this, separate datasets will be created for train, test, and validation datasets with genre classifications separated into separate rows. Predictive model exploration will include training on both sets of data.

Original data prior to cleaning

```
##    userId movieId rating timestamp                         title
## 1:      1     122      5 838985046               Boomerang (1992)
## 2:      1     185      5 838983525                Net, The (1995)
## 3:      1     292      5 838983421                Outbreak (1995)
## 4:      1     316      5 838983392                Stargate (1994)
## 5:      1     329      5 838983392 Star Trek: Generations (1994)
## 6:      1     355      5 838984474       Flintstones, The (1994)
##                            genres
## 1:                Comedy|Romance
## 2:          Action|Crime|Thriller
## 3:  Action|Drama|Sci-Fi|Thriller
## 4:       Action|Adventure|Sci-Fi
## 5: Action|Adventure|Drama|Sci-Fi
## 6:       Children|Comedy|Fantasy
```

Wrangled dataset with combined genres This method results in a dataset with the same number of rows as the original.

userId	movieId	rating	movieTitle	movieYear	genres	rateYear	rateMonth	rateHour
1	122	5	Boomerang	1992	Comedy|Romance	1996	08	06
1	292	5	Outbreak	1995	Action|Drama|Sci-Fi|Thriller	1996	08	05
1	316	5	Stargate	1994	Action|Adventure|Sci-Fi	1996	08	05
1	329	5	Star Trek: Generations	1994	Action|Adventure|Drama|Sci-Fi	1996	08	05
1	355	5	Flintstones, The	1994	Children|Comedy|Fantasy	1996	08	06
1	356	5	Forrest Gump	1994	Comedy|Drama|Romance|War	1996	08	06

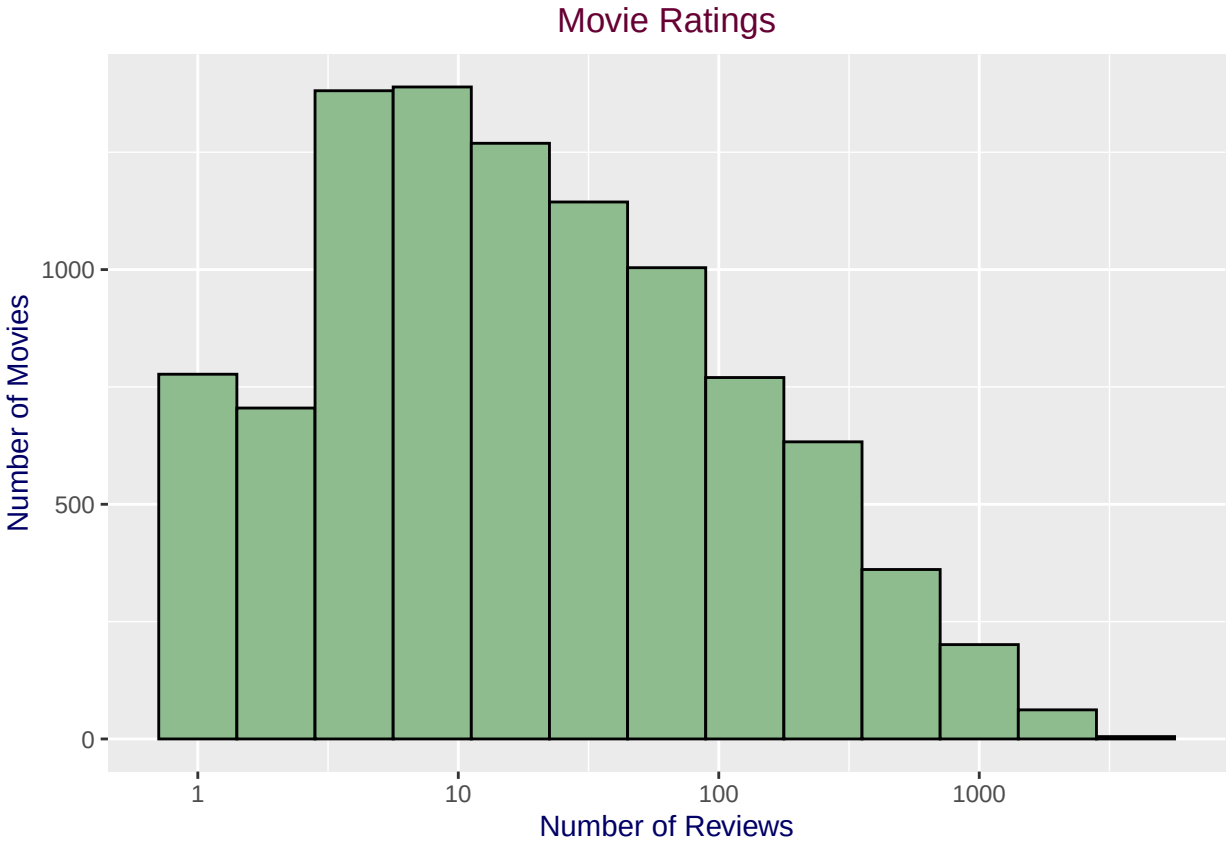
Wrangled dataset with genres separated into rows This method duplicates userID/movieId combinations increasing the total number of rows in the dataset.

userId	movieId	rating	movieTitle	movieYear	genres	rateYear	rateMonth	rateHour
1	122	5	Boomerang	1992	Comedy	1996	08	06
1	122	5	Boomerang	1992	Romance	1996	08	06
1	292	5	Outbreak	1995	Action	1996	08	05
1	292	5	Outbreak	1995	Drama	1996	08	05
1	292	5	Outbreak	1995	Sci-Fi	1996	08	05
1	292	5	Outbreak	1995	Thriller	1996	08	05

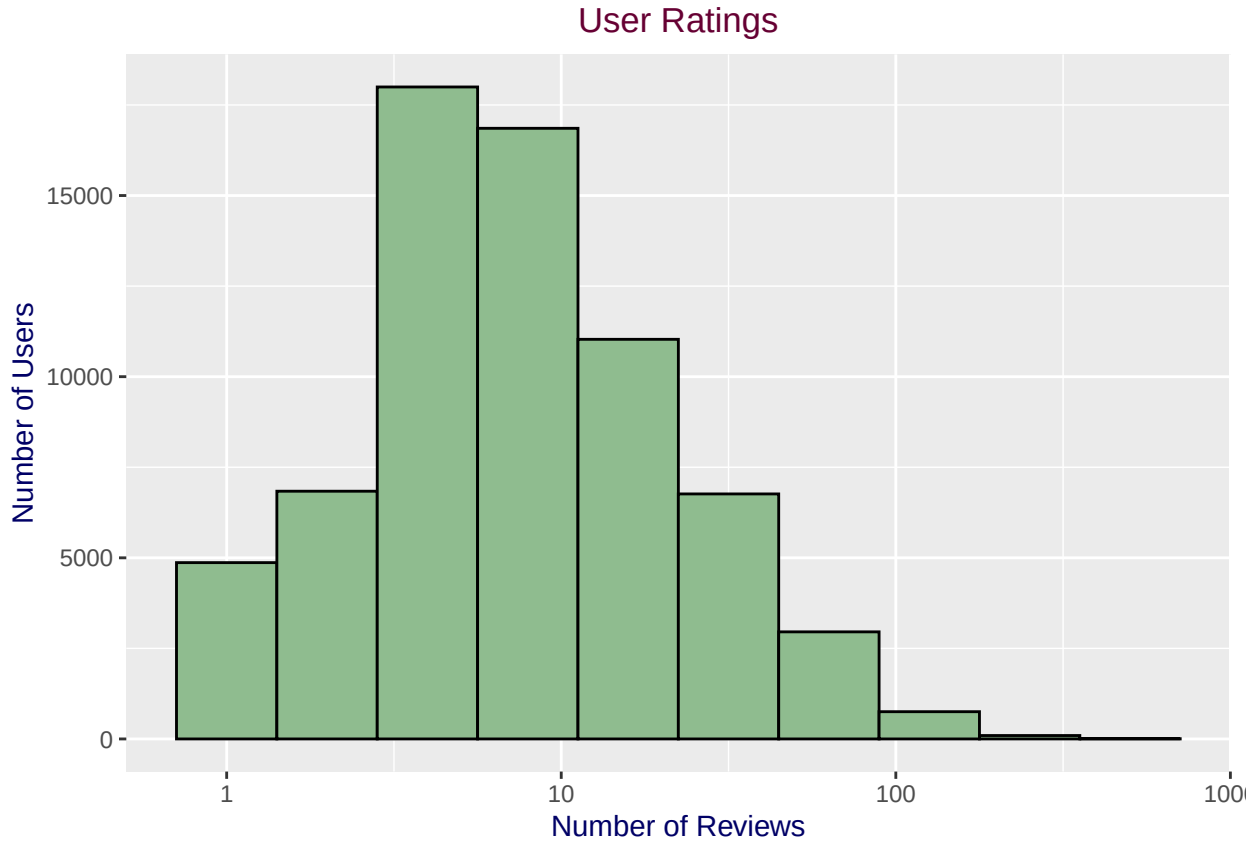
2.2 Data Exploration and Visualization

A review of the rating column shows there are 10 different rating scores between 0.5 and 5 with the median at 4 and mean at 3.513. From this we see the random chance of predicting the correct rating is 10%. Additionally, there are 9701 movies rated by 68159 users.

2.2a The Movies Further exploration of ratings shows that many movies are rated fewer than 5 times and only some movies are rated over 100 times. Because many movies have few ratings, regularization should be incorporated to penalize estimates that are formed using small samples.

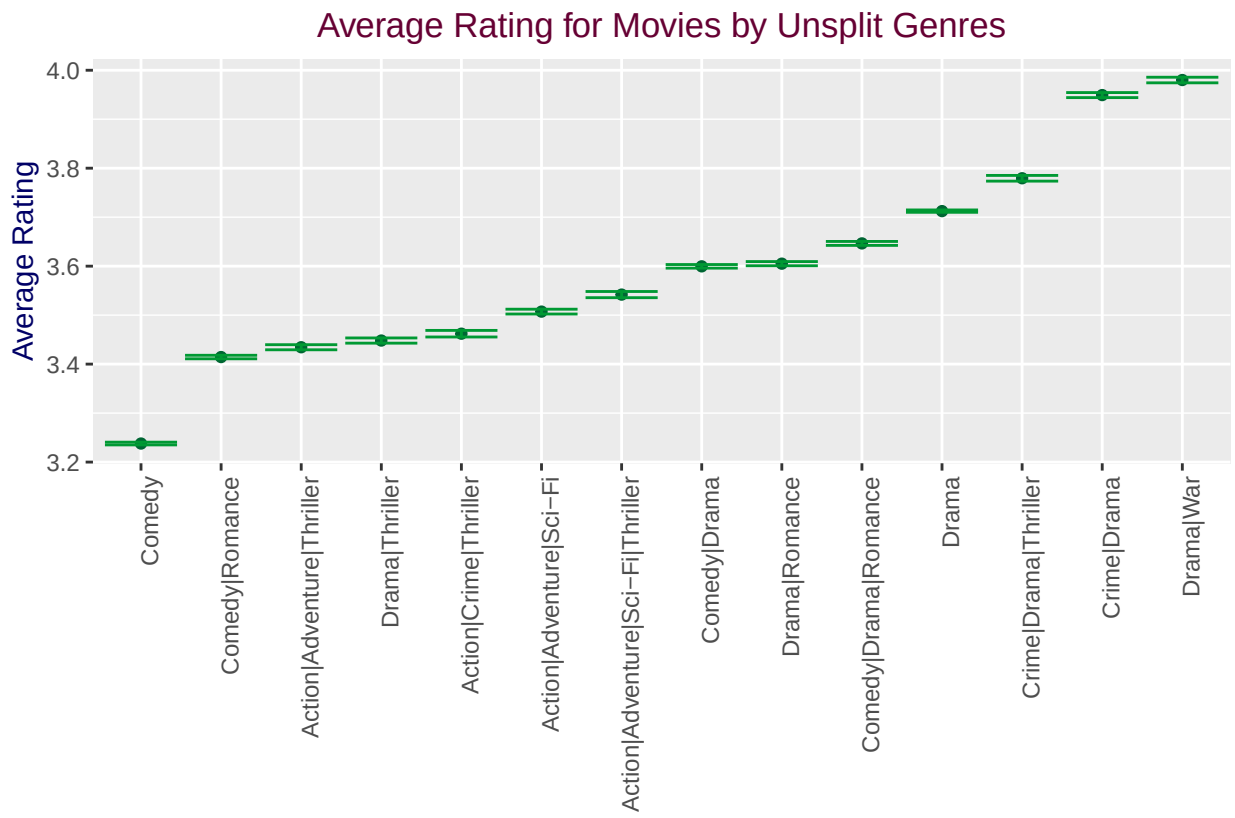


2.2b The Users A review of user rating behavior shows most users rate fewer than 100 movies and the majority rate fewer than 10. Regularization should be considered in user effect to account for many users giving few ratings.

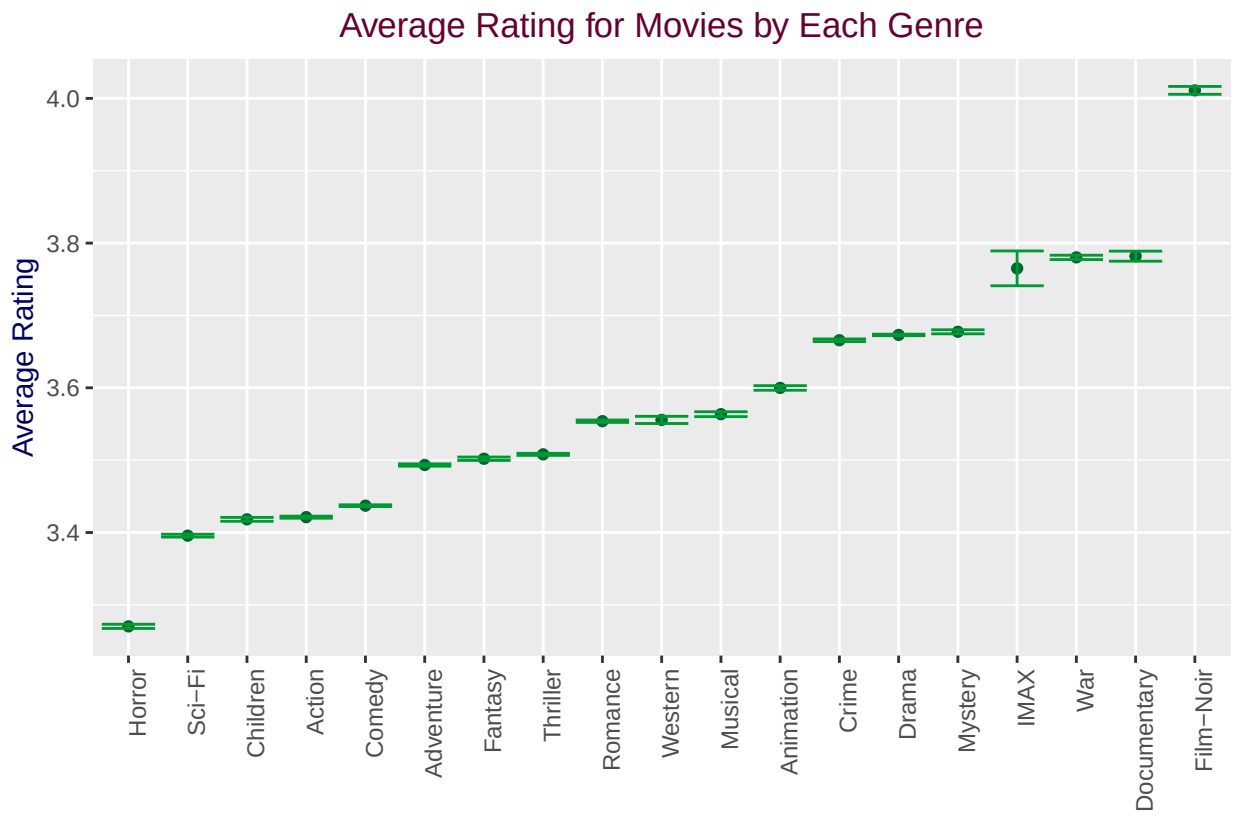


2.2c The Genre Another potential predictor is the genre of the movie. The following plots show evidence of differences in ratings based on genre. The first plot indicates genre categories that include *drama* are rated the highest. However, when genre categories are split, *drama* only ranks sixth among 20 classifications. Modeling on potential predictors should include analysis with genre categories as well as individual genre classifications.

Plot of unsplit genres

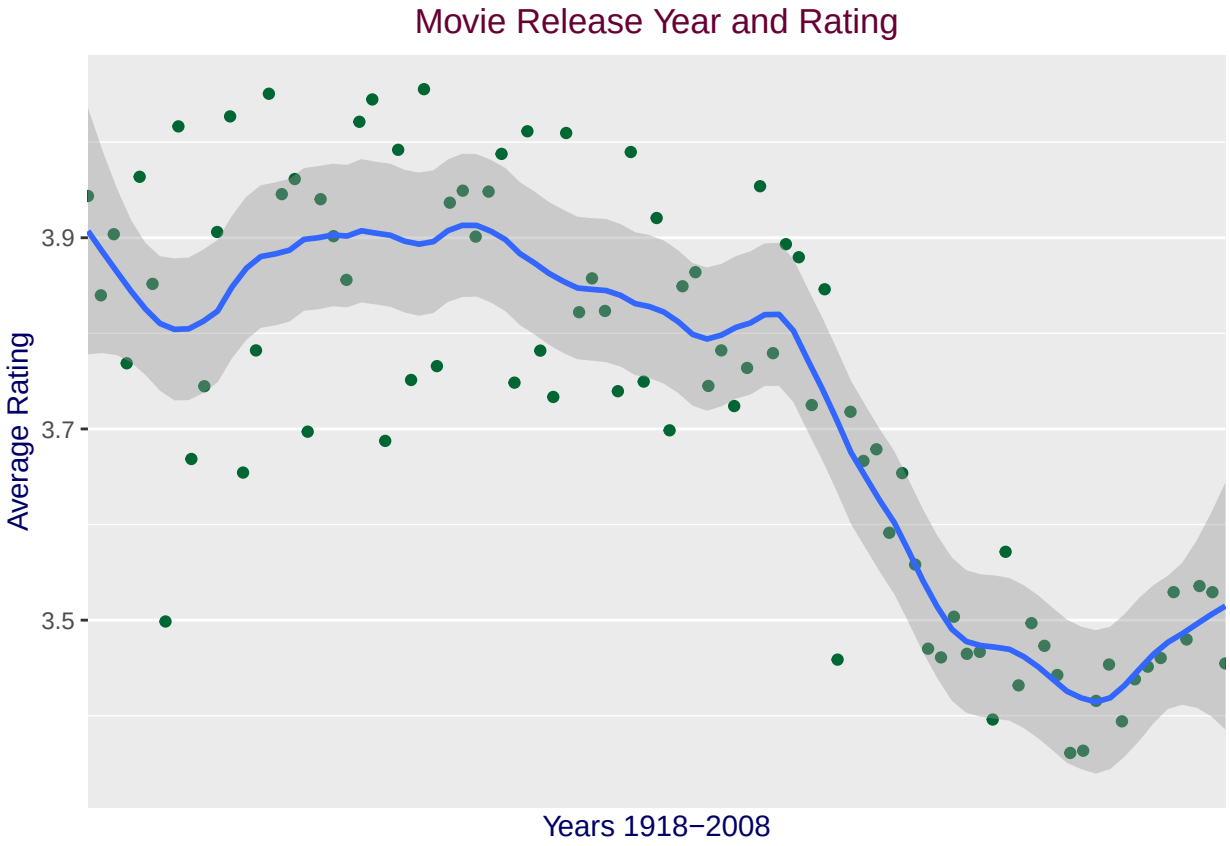


Plot of split genres

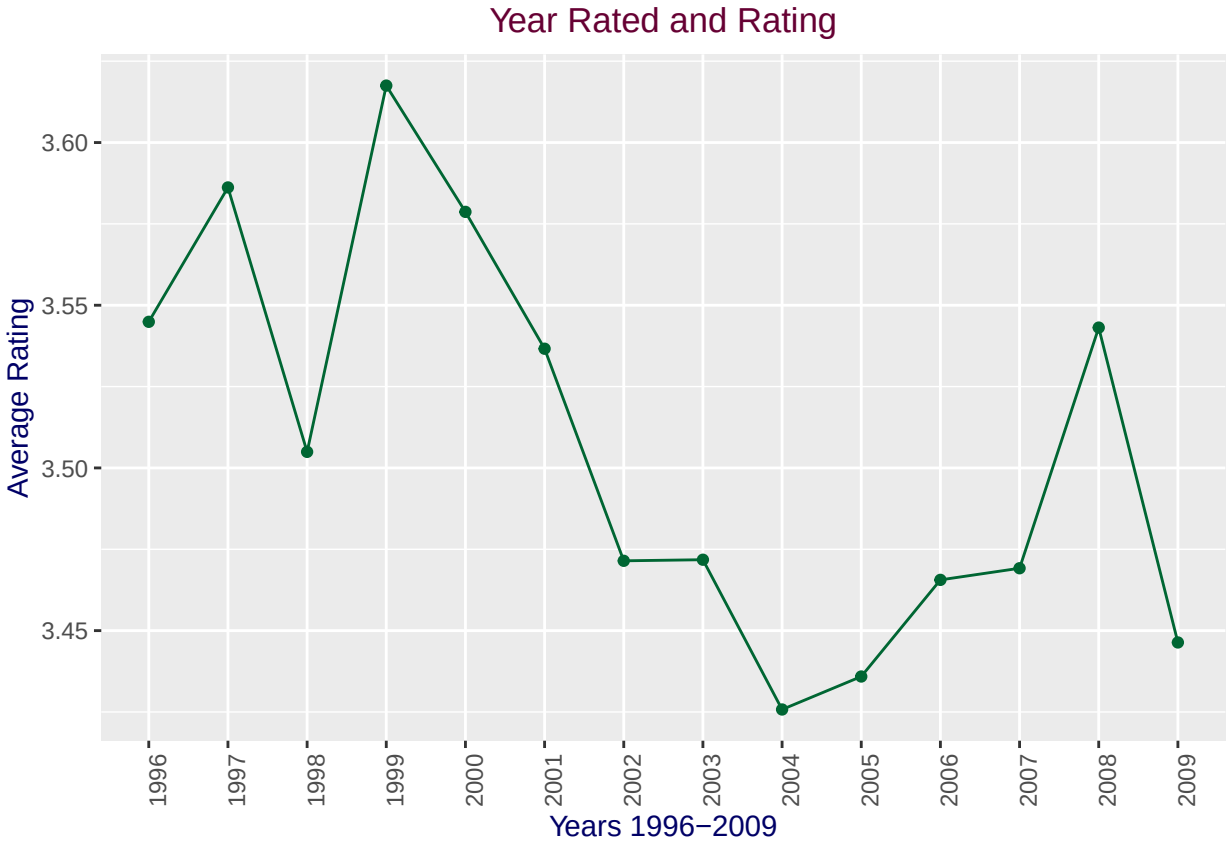


2.2d Other Potential Predictors This analysis also includes other potential predictors not included in the final predictive model. These are discussed in section 4.3, Recommendations for Future Study.

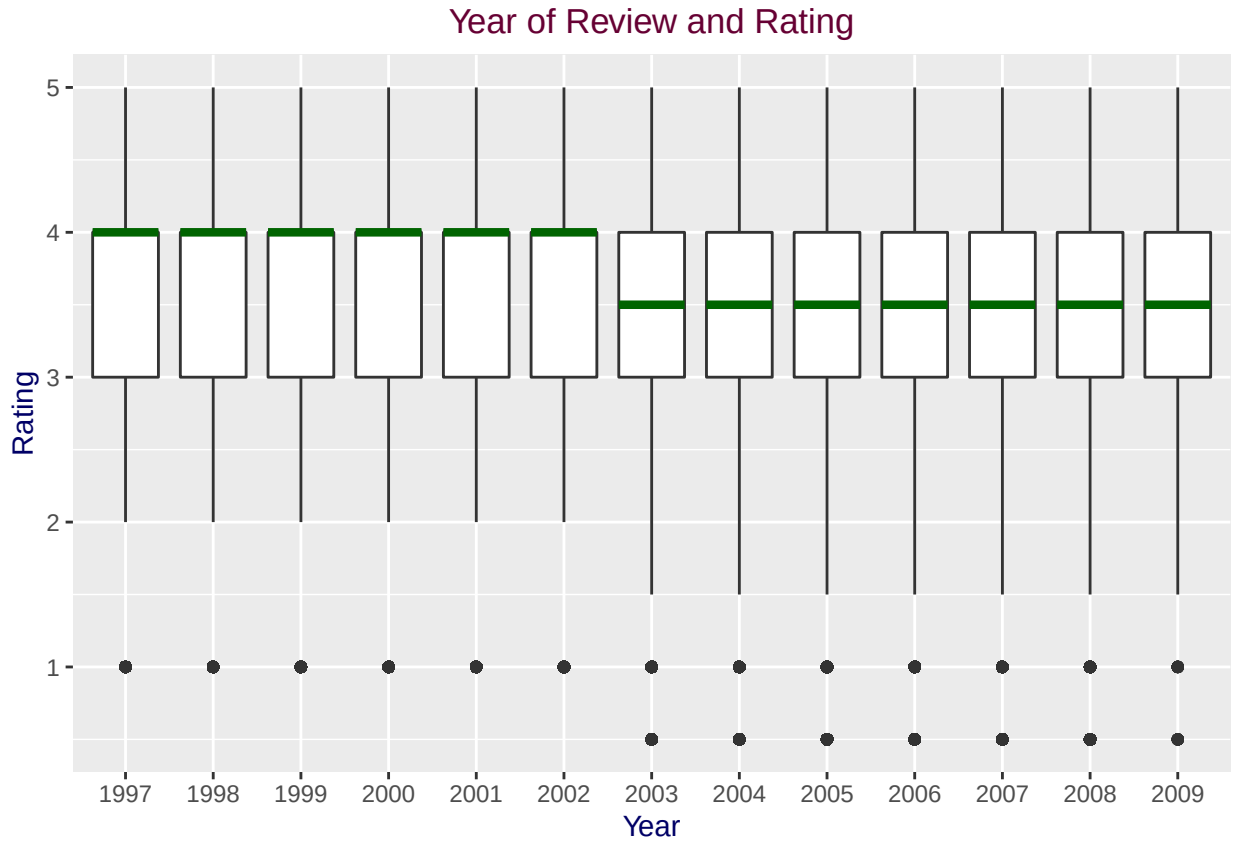
The following plot suggests a relationship between the year a movie was released and rating. Ratings appear to decrease over years. Loess smoothing was included to show a possible multiple regression approach for movie release year as a predictor. There seems to be predictive power in the year the movie was released, except the conditional probability is not linear.



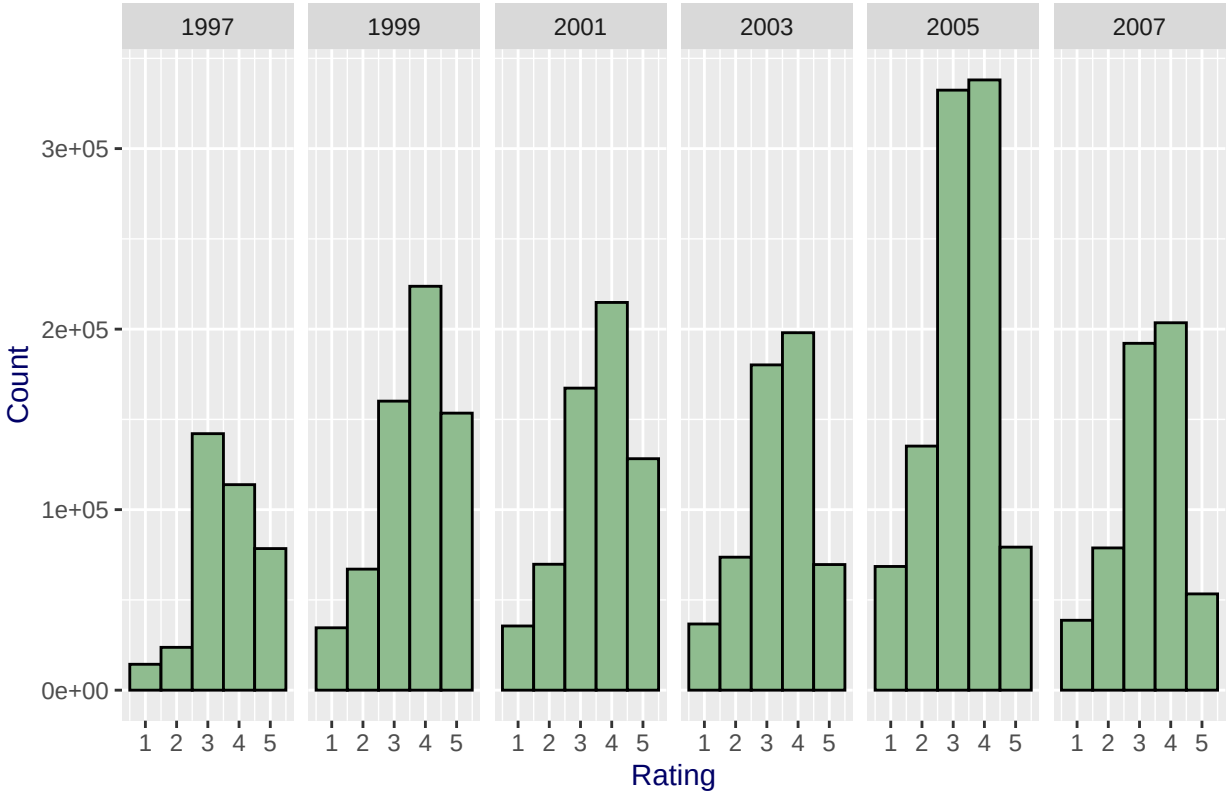
There also seems to be a relationship between the year a movie was rated and the rating. Similar to the year a movie was released, overall ratings decreasing over years. For 1996-2009, the lowest average rating was in 2004 with a rating of 3.426 and the highest was in 1999 with a rating of 3.618.



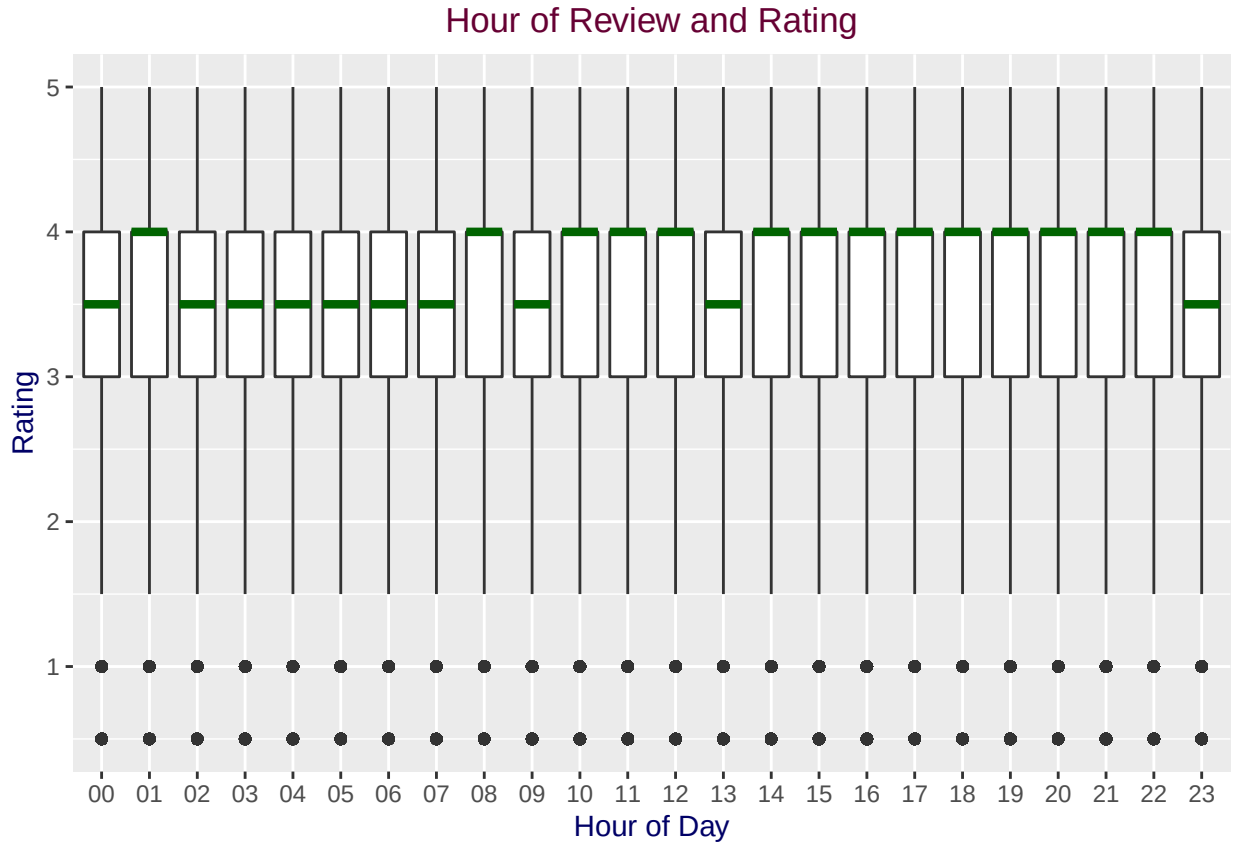
A boxplot of the year of ratings show equal boxes for all years. Although, median rates, highlighted in green, are at 75% for 1997-2002.



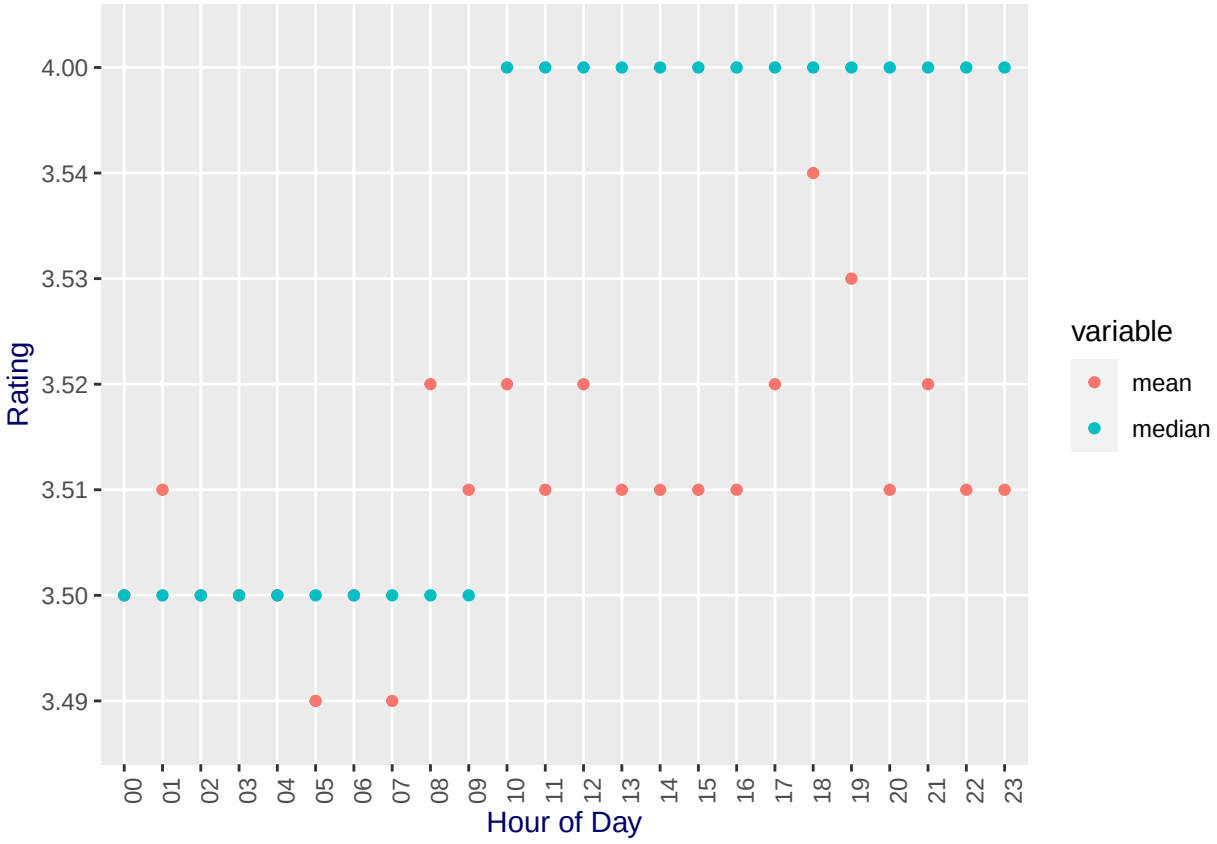
Further exploration of ratings by year of review show a left skewed distribution for years before 2003. For predictive analysis, a normalizing transformation might be considered.



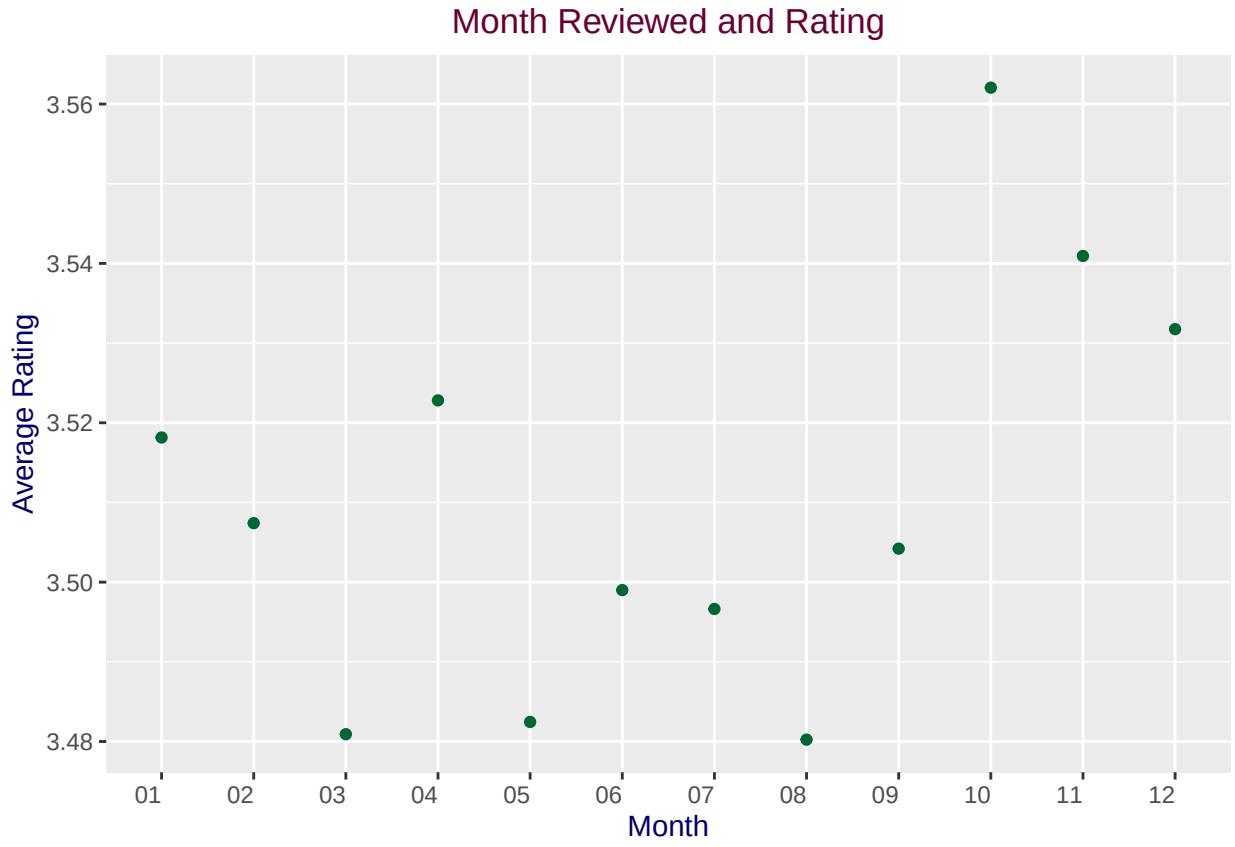
The following boxplot explores the hour of day and differences in ratings. The middle 50% of all boxplots from a sample of the train set are quite similar. However, median ratings remain consistent for several hours, fluctuating between 3.5 and 4.



Further exploration on hour of day as a potential predictor shows a plot with mean and median ratings by hour of day. In most instances the mean is lower than the median. This indicates a left or negatively skewed data. The most common values in the distribution might not be near the mean. Additionally, this skewed data can affect which types of analyses to perform.



One final plot is included to show that the month of rating may be a potential predictor. This plot shows higher average ratings for months 10-12.



2.3 Analysis of the Data

A standard way to measure the error in a predictive model is with Root Mean Square Error (RMSE). This is the function to calculate RMSE that will be used for predictive analysis.

```
RMSE <- function(true_ratings, predicted_ratings){
  sqrt(mean((true_ratings - predicted_ratings)^2))
}
```

Naive Baseline Model The first step to prediction is to calculate the average rating of all movies(mu). From the train dataset, the mean of ratings is 3.512.

Naive Mean-Baseline Model Our next approach is to apply our RMSE formula to mu. The formula used is:

$$Y_{u,i} = \hat{\mu} + \varepsilon_{u,i}$$

with $\hat{\mu}$ as the mean(mu) and $\varepsilon_{u,i}$ as independent errors centered at 0.

The RMSE using the test dataset with this model is 1.06. Since this is result is above the target RMSE of < 0.86490, further modeling needs to be applied.

Adding Movie Effect The next predictor considers the effect of individual movies; some movies are rated higher than others. The addition of b_i accounts for movie effect with this formula:

$$Y_{u,i} = \hat{\mu} + b_i + \epsilon_{u,i}$$

The RMSE result using the test dataset on the movie effect model is 0.943. Since this result is still above the target RMSE of < 0.86490, further predictors are considered.

Adding User Effect Based on the user/ratings plot, there is a user effect. Users are all different and rate movies differently. Our next model adds movie effect(bu) to our model with the following formula:

$$Y_{u,i} = \hat{\mu} + b_i + b_u + \epsilon_{u,i}$$

The RMSE result using the test dataset on the user effect model is 0.865. This is result is getting better, but still above the target RMSE so further predictors are added.

Adding Genre Effect Previous plots demonstrate a potential genre effect on movie predictions. Since the original dataset (edx) included combined genres, calculations are performed on combined genres as well as each genre. Additionally, mu (mean of the train set) was used to compare with mu_g (mean of train_genre set) due to increased number of rows in train_genre table.

The mean of the original train set(mu) is 3.512 and the mean of the genre training set that is split into separate rows is 3.527

This model adds the genre effect(bg) to our previous model:

$$Y_{u,i} = \hat{\mu} + b_i + b_u + b_g + \epsilon_{u,i}$$

The RMSE result using the test dataset with mu on the genre effect model is 0.863. Using the rating average on the genre set with separated rows with mu_g is also 0.863. Continued modeling will include both approaches to detect any variability in regularization.

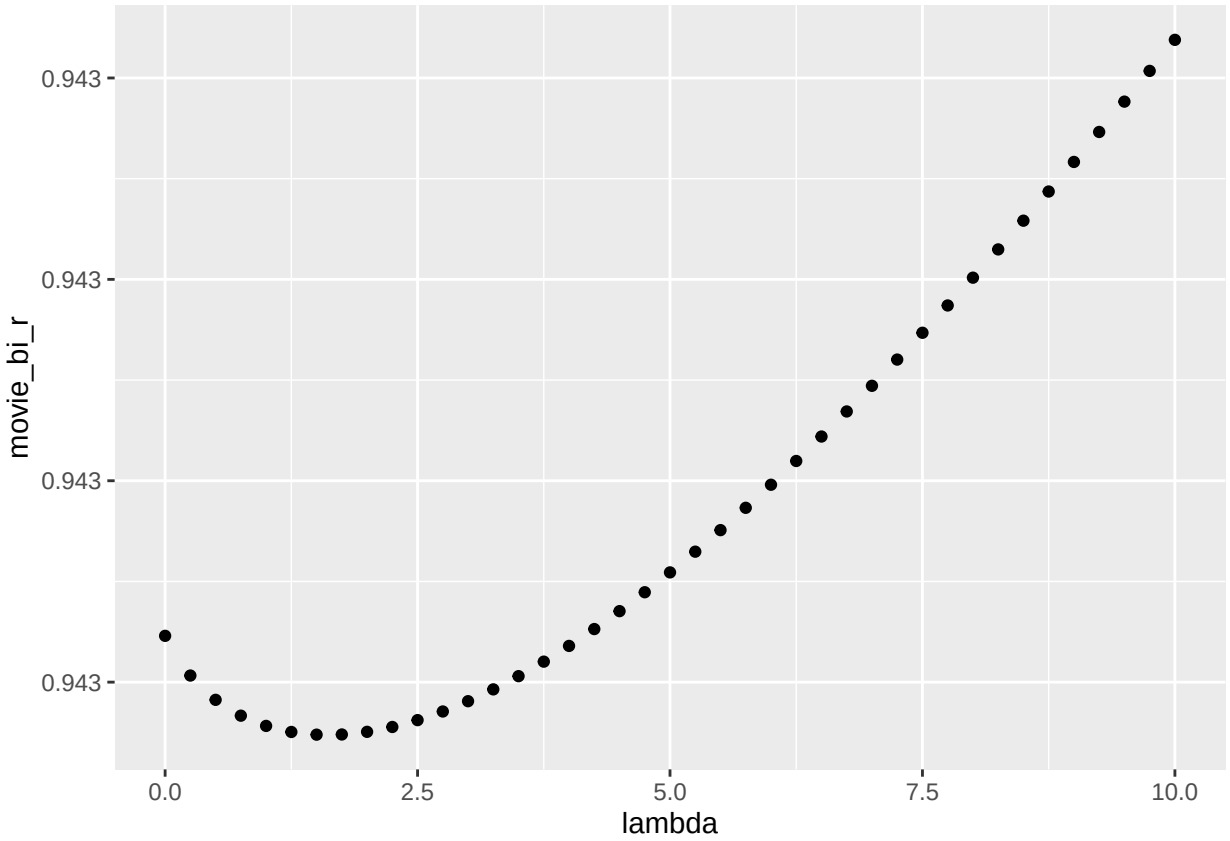
Regularization Since many ratings in the data set come from a small sample size for all three predictors, a penalty (lambda) will be added. Regularization will be used to limit the total variability of effect size.

The formula for movie effect penalty includes n, which is the number of ratings made for movie i. When the number of ratings for a movie is large, the penalty is negligible.

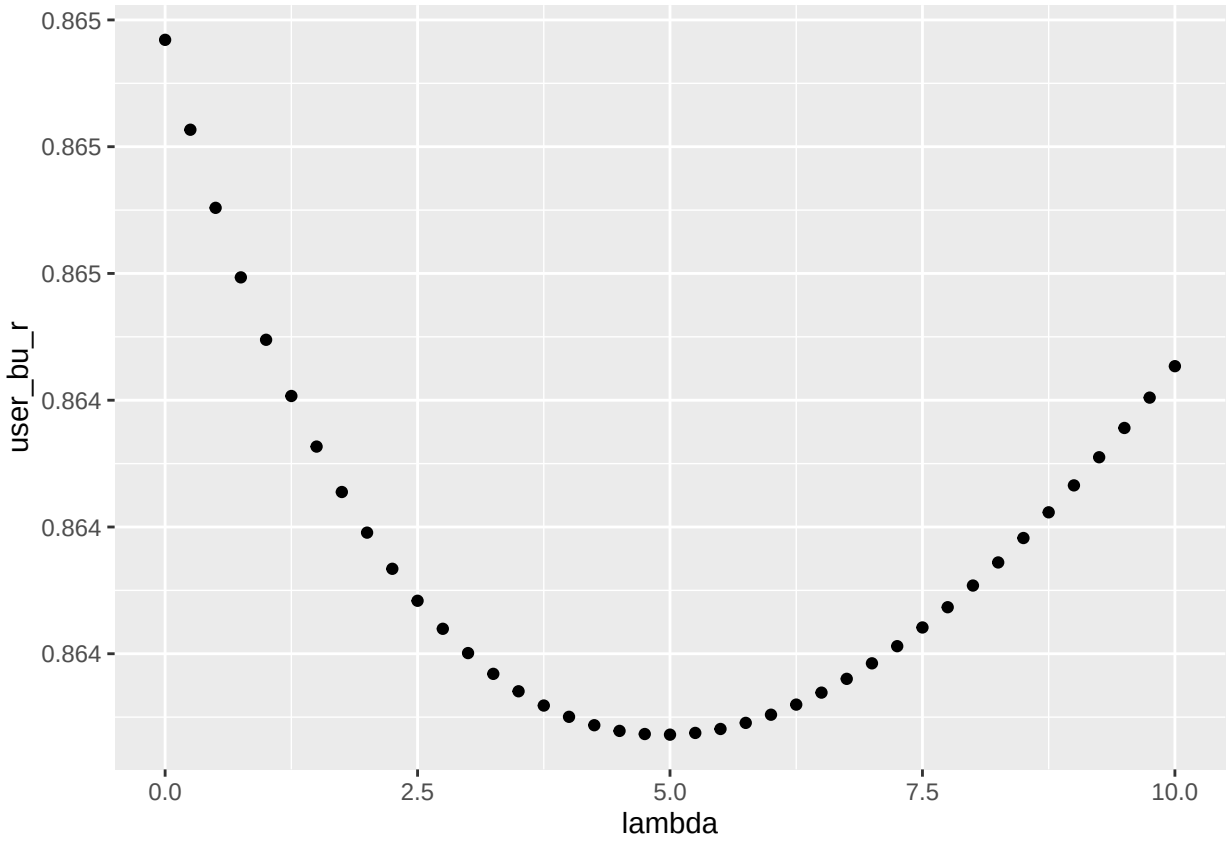
$$\hat{b_i}(\lambda) = \frac{1}{\lambda + n_i} \sum_{u=1}^{n_i} (Y_{u,i} - \hat{\mu})$$

The penalty(lambda) was selected using cross-validation from the series of possibilities including numbers 0 to 10 with divisions at .25. This series produced acceptable tuning parameters for all three models.

The first model, using movie effect as the predictor, returned a lambda value of 1.5 with the updated RSME at 0.943. Visualizing this also shows that the selected series of possibilities includes the minimum value for lambda.

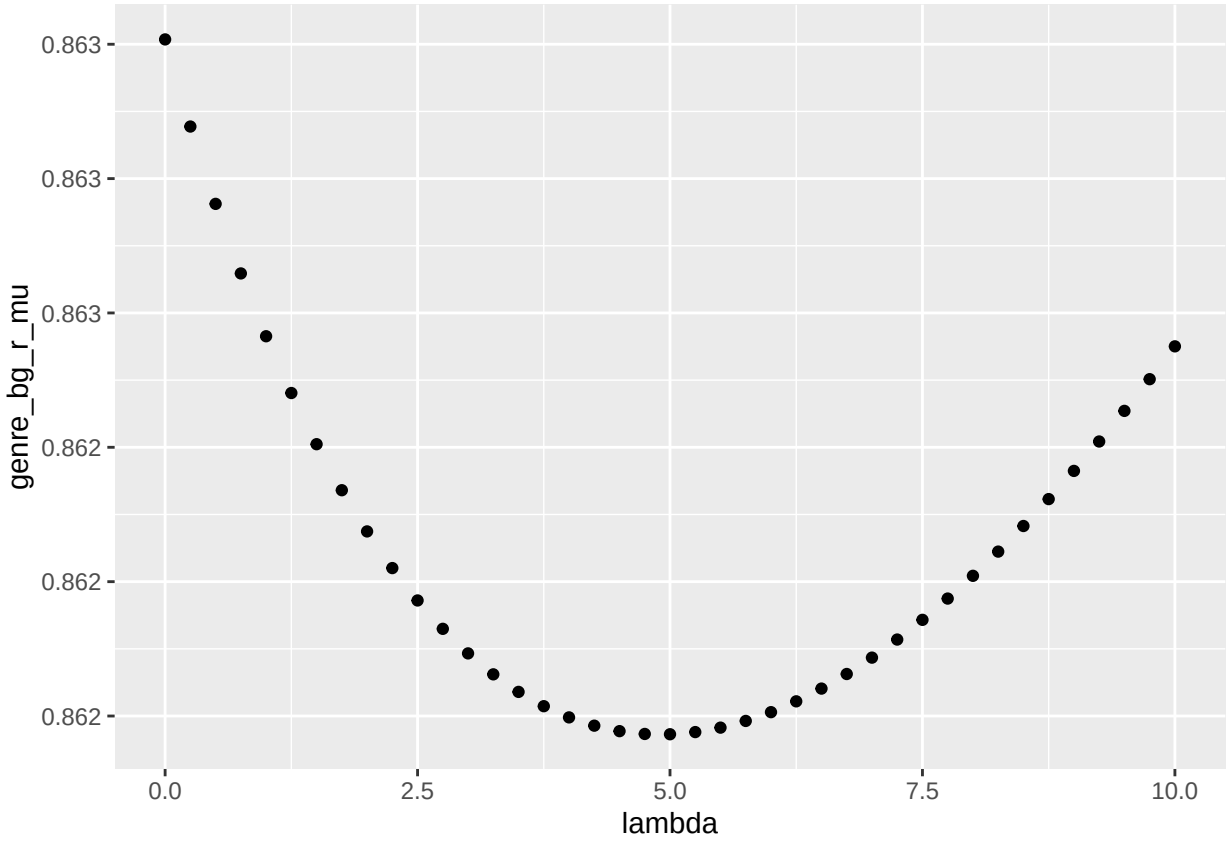


The second model, adding user effect, returned a lambda value of 5 with the RMSE user effect adding regularization at 0.864. Visualizing this also shows that the selected series of possibilities includes the minimum value of lambda for user effect.

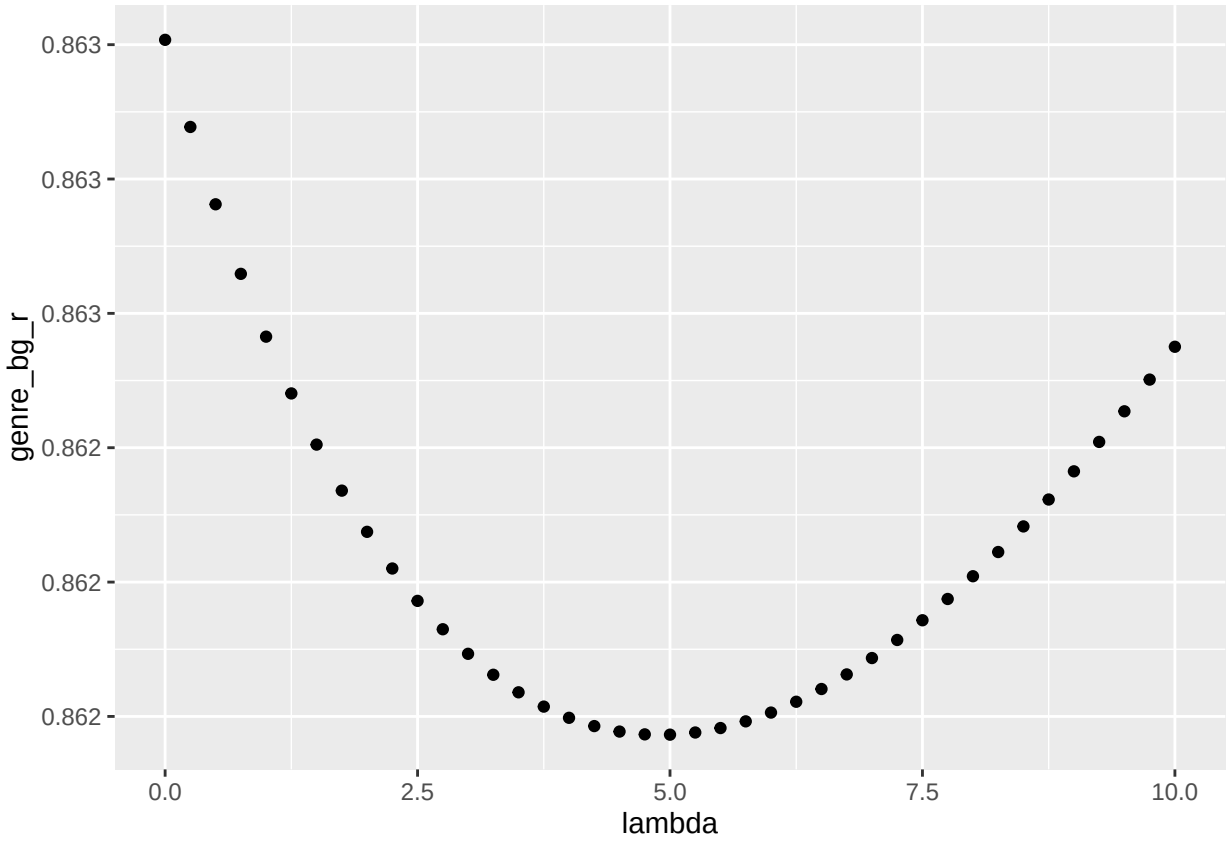


The final model, which adds genre effect with regularization, was determined by comparing RSME results for average rating scores of mu (average rating) of the train set and mu_g (average rating) of the train_genre set.

The model using genre effect with mu returned a lambda value of 5 with the RMSE regularization at 0.862. Visualizing this also shows that the selected series of possibilities includes the minimum value of lambda for user effect.



The model using genre effect with mu_g, average rating of the genre train set, returned a lambda value of 5. Visualizing this also shows that the selected series of possibilities includes the minimum value of lambda for user effect.



The RMSE regularization with this model is 0.862. When adding additional decimal places, this result is slightly better than mu genre effect and will be used for the final model.

Method	RMSE
Regularized Movie+User+Genre With mu	0.862186434411
Regularized Movie+User+Genre With mu_g	0.862186434348

3. Results

The final step is evaluating the RMSE of the final algorithm to the final hold-out, the validation dataset. This is the only time the validation set has been used in this project. All prior RMSE scores were obtained using the train and test datasets. As the final RMSE score shows, the final model exceeds expectations of RMSE < 0.86490 for this movie recommendation model.

Conclusion	RMSE
Final Model Used On Validation Set	**0.86405**

4. Conclusion

4.1 Summary of Findings

This study successfully trained a machine learning algorithm to predict movie ratings with a final RMSE value of 0.86405. Predictions used for this model included three predictors: movie effect, user effect, and genre effect. The movie effect indicated that some movies are better than others and receive higher ratings, and others are worse and get lower reviews. User effect showed that some users give higher ratings than others and genre effect demonstrated that some genres receive higher reviews than others.

Regularization was used on all three predictors to account for differences in sample size. This approach was successful as RMSE results were lower when regularization was applied to all predictors.

The following table shows RMSE results modeled with the train set and substantiated with the test set.

Test_Prediction	RMSE
Basic average	1.060053702224
Movie effect	0.942961498005
User effect	0.864684294902
Genre mu effect	0.862703593210
Genre mu_g effect	0.862703593210
Regularized movie effect	0.942936965846
Regularized user effect	0.864136179290
Regularized mu genre effect	0.862186434411
Regularized mu_g genre effect	0.862186434348

From here, the final model was applied to the validation set where the RMSE result confirms a successful movie prediction algorithm.

Validation_Prediction	RMSE
Final Model Used On Validation Set	**0.86405450983**

4.2 Limitations

One challenge with this study was the size of the dataset. Many processes were slow and better lambda values may have been found through further calculations. Also, including additional predictors to the dataset would have been interesting (e.g. age of user, gender of user, producer of movie, money spent to make the movie).

4.3 Recommendations for Future Study

Future studies may consider potential predictors from the year movie was released, year movie was rated, hour of day the movie was rated, and month of year the movie was rated. Section 2.2 of this paper, Data Exploration and Visualization, included promising visual analysis of these potential predictors that weren't included in this model prediction.