MovieLens Capstone Project

John Sowers

9/14/2020

INTRODUCTION

This project report is submitted to fulfill part of the requirement for the Edx Data Science: Capstone Course, utilizing the Movielens data set downloaded from the GroupLens website at http://files.grouplens.org/datasets/movielens/ml-10m.zip. The program code for this project was compiled utilizing R version 4.02 / R Studio version 1.3.1073. It was run on a Dell m15 R3 laptop computer with an i7 Central Processing Unit (CPU) with 16 Gigabytes of Random Access Memory (RAM) utilizing Microsoft Windows 10.

The basis of this project is taken from the 2006 Netflix challenge in attempts to increase its movie rating prediction accuracy, using the Residual Mean Square Error (RMSE) value to evaluate the results. This original data set was broken into 3 data sets: (1) the training data consisting of 100,480,507 reviews; (2) the quiz data consisting of 1,408,342 reviews; and (3) the test data consisting of 1,408,789 reviews.

In this project, the smaller Movielens data set is also separated into 3 distinct data sets: (1) the training data consisting of 8,100,054 reviews; (2) the test data consisting of 900,007 reviews; and (3) the validation data consisting of 999,993 reviews. The training and test sets are utilized to analyze, create, and tune the models with the validation data reserved to evaluate the final tuned model. The structure of the combined training and test data is shown below.

```
## Classes 'data.table' and 'data.frame':   9000061 obs. of  6 variables:
##  $ userId   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ movieId  : num  122 185 231 292 316 329 355 356 362 364 ...
##  $ rating   : num  5 5 5 5 5 5 5 5 5 5 ...
##  $ timestamp: int  838985046 838983525 838983392 838983421 838983392 838983392 ...
##  $ title    : chr  "Boomerang (1992)" "Net, The (1995)" "Dumb & Dumber (1994)" ...
##  $ genres   : chr  "Comedy|Romance" "Action|Crime|Thriller" "Comedy"  ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

Analysis of the combined training and test data set reveals that submitted reviews cover 10,677 unique movies in 20 distinct genres that were released over a 94-year period by 69,878 separate users during a course of 15 years.

METHODS / ANALYSIS

As with the original Netflix challenge, the RMSE is used again to evaluate the results in this project. The RMSE evaluation method is chosen due to its ability to manage large databases with multiple predictors. The RMSE equation is shown below.

$RMSE = \sqrt{\sum_{i=1}^{n}(\hat{y} - y)^2/n}$

In this equation, the variable n denotes the total number of reviews, $\hat{y}$ denotes the movie's true rating, and y denotes the movie's predicted rating that the model returns.

Unlike the original Netflix challenge with teams using a much large data set and many predictors to obtain low RMSE values, the scope of this paper uses a smaller number of reviews with only 6 parameters to to reduce the RMSE below 0.86490. These parameters include:

```
## [1] "userId"    "movieId"   "rating"    "timestamp" "title"     "genres"
```

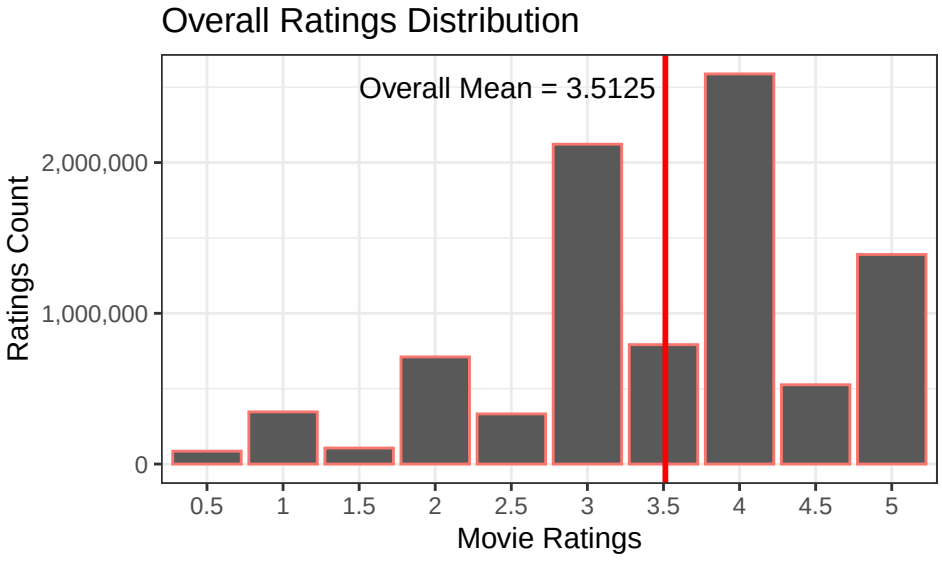
From these 6 parameters, five predictors are identified for further analysis and model creation. These models include:

Model 1: $Y = \mu + \epsilon$
Model 2: $Y_u = \mu + b_i + \epsilon_u$
Model 3: $Y_{i,u} = \mu + b_i + b_u + \epsilon_{i,u}$
Model 4: $Y_{u,u,g} = \mu + b_i + b_u + b_g + \epsilon_{i,u,g}$
Model 5: $Y_{i,u,g,t} = \mu + b_i + b_u + b_g + b_t + \epsilon_{i,u,g,t}$

With the exception of Model 1 that uses the overall average as the sole basis for its prediction, these models benefit from regularization with a variable λ introduced into the models. This helps reduce, but not eliminate, the effects of outliers on the RMSE values and prevent overfitting. These models require further tuning to find the optimal λs that minimize the RMSE values based on the test data. These tuned models are ran against the validation data to determine each model's actual performance.

Model 1 (μ)

The first model utilizes the overall mean of the movie ratings as the sole predictor. Analyzing the data, the mean of the overall ratings is 3.5124 and the median is 4.0; and this can be seen with 16% of the ratings falling between 0 and 3 and 74% of the ratings ranging between 3 and 5. Additionally, 80% of the ratings comprised of whole numbers and 20% comprised of non-whole numbers. This initial analysis is shown in the below figure.



Based on these observations, a first simple model is constructed, as depicted in the equation:

$Y = \mu + \epsilon$

In this equation, the variable Y denotes the model's predicted rating, μ denotes the total overall mean, and ϵ denotes the residual error. No tuning of the model is required for this basic model. Utilizing the test data results in the following output.

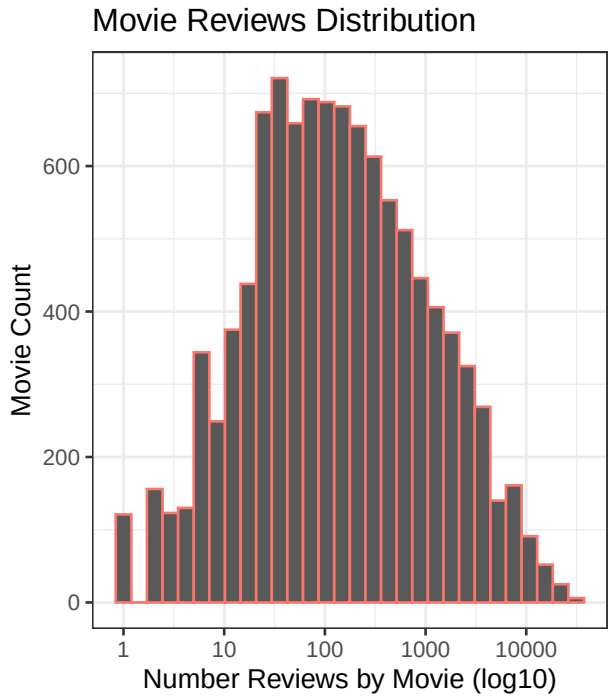
```
##    Models Predictors Lambda  RMSE
## 1 Model 1         mu      0 1.059
```

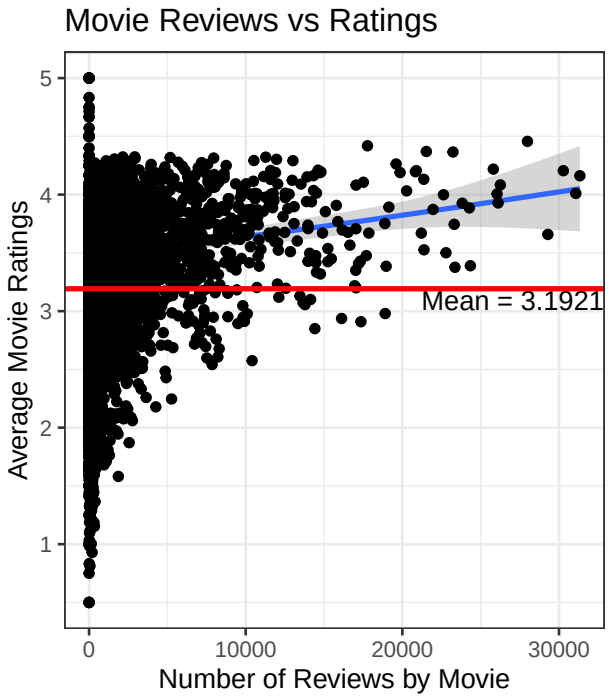
While the result of Model 1 is far from the goal set by this project, it provides a basis to compare future models.

Model 2 (μ + Movie Effects)

Model 2 adds movie effects to that of μ, as described in Model 1.

The number of movie reviews submitted for movies range from a low of 1 to a high of 31,336. Likewise, ratings range from a minimum score of 0.5 and a maximum of 5.0 with the mean rating of distinct movies at 3.1921. This mean movie rating is lower than the mean of the overall ratings of 3.5125. Movie reviews and ratings distribution are depicted in the following figures.





Outliers exist from movies that have few ratings, negatively impacting the RMSE calculations. Such outliers can be seen by inspecting the number of reviews of the highest and lowest rated movies as seen in the following tables:

```
## # A tibble: 5 x 4
##   movieId num_reviews avg_rating title
##     <dbl>       <int>      <dbl> <chr>
## 1    3226           1          5 Hellhounds on My Trail (1999)
## 2   33264           2          5 Satan's Tango (SÃ¡tÃ¡ntangÃ³) (1994)
## 3   42783           1          5 Shadows of Forgotten Ancestors (1964)
## 4   51209           1          5 Fighting Elegy (Kenka erejii) (1966)
## 5   53355           1          5 Sun Alley (Sonnenallee) (1999)
```

```
## # A tibble: 5 x 4
##   movieId num_reviews avg_rating title
##     <dbl>       <int>      <dbl> <chr>
## 1   64999           2       0.75 War of the Worlds 2: The Next Wave (2008)
## 2    5805           2       0.5  Besotted (2001)
## 3    8394           1       0.5  Hi-Line, The (1999)
## 4    8707           1       0.5  Grief (1993)
## 5   61768           1       0.5  Accused (Anklaget) (2005)
```

Based on these observations, a model is constructed, expanding on Model 1 that introduces the variable b_i. This new variable accounts for movie effects while λ reduces, but not eliminates, the impact of the outliers by movies with few reviews. This equation is written as follows:

$Y_i = \mu + \lambda_i b_i + \epsilon_i$

This model returns the following outputs when it is ran against the test data with the first result depicting the model without regularization (i.e. $\lambda = 0$) and the second result depicting the model results with regularization (i.e. λ set to that of its best tuned value).

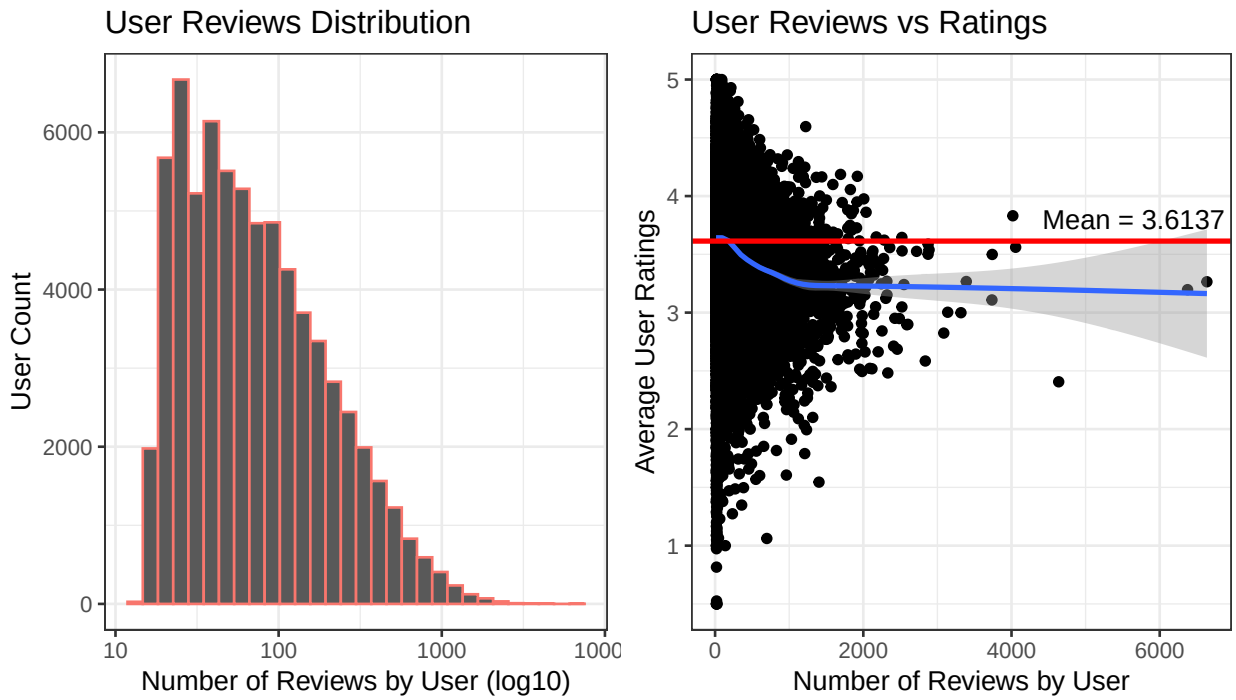
```
##    Models   Predictors Lambda    RMSE
## 2 Model 2     mu+Movie   0.00 0.94266
## 3 Model 2 mu+Movie+Reg   2.25 0.94261
```

Comparing the unregularized result with that of the Model 1 result shows that this model provides a RMSE reduction of 0.11634 below that of the result from Model 1. Additionally, Model 2's regularized result reduces the outlier effects by an additional 0.00005.

Model 3 (μ + Movie + User Effects)

Model 3 adds user effects to that of the μ and movie effects that were described in Model 2.

Reviews from distinct users range from a low of 13 to a high of 6,637. Likewise mean user ratings range from a minimum of 0.5 to a maximum of 5.0 with the mean average user rating of 3.6137. This mean user rating is higher than the mean of the overall ratings of 3.5125. User reviews and ratings' distribution are depicted in the following figures.



Based on these observations, a model is constructed, expanding on Model 2 and introducing the variable b_u. This new variable accounts for user effects while λ is used to reduce the impact of the outliers by users with few reviews. This equation is written as follows:

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

This model returns the following outputs when it is ran against the test data with the first result depicting the model without regularization (i.e. $\lambda = 0$) and the second result depicting the model with regularization (i.e. λ set to that of its best tuned value).

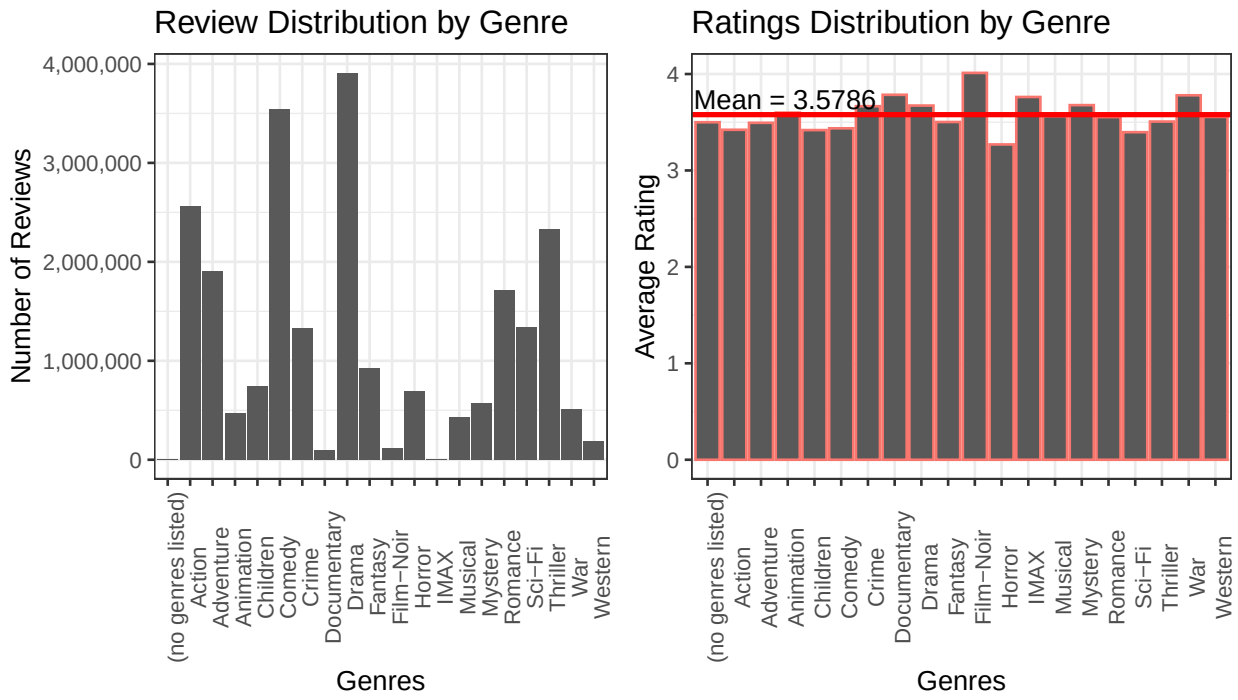
```
##    Models        Predictors Lambda    RMSE
## 4 Model 3     mu+Movie+User      0 0.86461
## 5 Model 3 mu+Movie+User+Reg      5 0.86401
```

Comparing the unregularized result Model 3 with that of Model 2 shows that there is a further RMSE reduction of 0.07799. Additionally, Model 3's regularized result reduces the outlier effects by an additional .0006. At this point, based on the test data from which the model was tuned, the RMSE is below that of the project goal of 0.8940; however, the final results are based on an independent validation data set, which will have higher RMSE values. Further analysis and model creation is warranted to meet the project's objective.

Model 4 (μ + Movie + User + Genre Effects)

Model 4 adds genre effects to that of the μ, movie effects, and user effects that were described in Model 3.

Although there are 20 distinct genres in the Movielens data set, movies may fall under multiple genres with the data containing 797 genre combinations. Reviews from distinct genres range from the low of 13 to a high of 6,637. Likewise the minimum mean genre rating is 1.47 with the maximum rating of 4.71. The overall genre mean rating is 3.5786 which is higher than the mean of the overall mean rating of 3.5125. Genre reviews and ratings distribution are depicted in the following figures.



Outliers exist from genres that have few ratings, which negatively impact a model's rating predictions. Such outliers include the genres "Film-Noir" with 118,394 reviews on 148 movies with an average rating of 4.0117 as well as the genre "IMAX" with 8,190 reviews on 29 movies with an average rating of 3.7618. Results of the top rated genres is shown in the following table.

```
##         genres users movies reviews ratings
## 1    Film-Noir 31332    148  118394  4.0117
## 2  Documentary 24325    481   93252  3.7844
## 3          War 64943    510  511330  3.7795
## 4         IMAX  6411     29    8190  3.7618
## 5      Mystery 61869    509  567865  3.6774
## 6        Drama 69863   5336 3909401  3.6730
## 7        Crime 68683   1117 1326917  3.6662
## 8    Animation 59006    286  467220  3.5996
## 9      Musical 58937    436  432960  3.5628
## 10     Western 47632    275  189234  3.5551
```

Based on these observations, a model is constructed, expanding on Model 3 and introducing the variable b_g. This new variable accounts for genre effects while λ is again used to reduce the impact of the outliers by genres with few reviews. This equation is written as follows:

$$Y_{i,u,g} = \mu + \lambda_{i,u,g} b_i + \lambda_{i,u,g} b_u + \lambda_{i,u,g} b_g + \epsilon_{i,ui,g}$$

This model returns the following outputs when it is ran against the test data with the first result depicting the model without regularization (i.e. $\lambda = 0$) and the second result depicting the model with regularization (i.e. λ set to that of its best tuned value).

```
##    Models              Predictors Lambda    RMSE
## 6 Model 4     mu+Movie+User+Genre      0 0.86426
## 7 Model 4 mu+Movie+User+Genre+Reg      5 0.86369
```

Comparing Model 4's unregularized result with that of Model 3 shows that there is a further RMSE reduction of 0.00035. Additionally, Model 4's regularized result reduces the outlier effects by an additional 0.00057. Again, these results meet the project goal based solely on the test data; however, the final results are to be ran against the separate validation data and further analysis of the time effect is warranted.

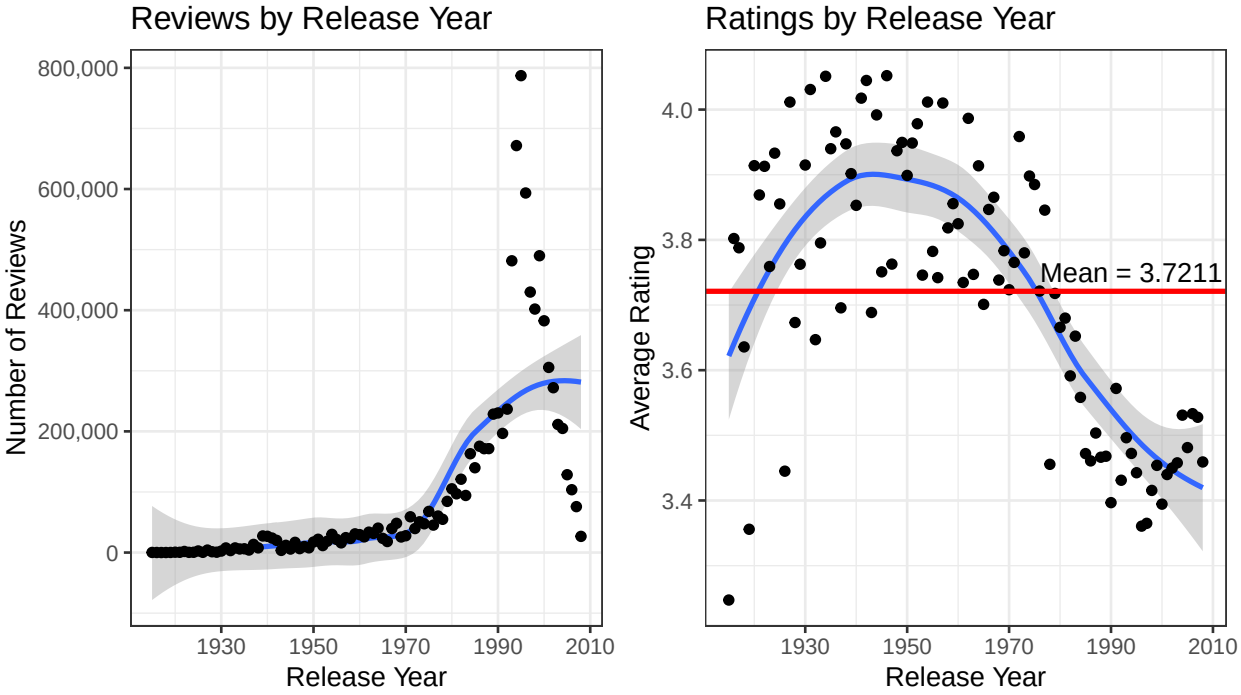
Model 5 (μ + Movie + User + Genre + Time Effects)

Model 5 adds the time effects to that of the μ, movie effects, user effects, and genre effects that were described in Model 4.

This final model compensates for the change of time as described in four distinct groups: (1) Year of Release; (2) Year of Review; (3) Week of Review; and (4) Hour of Review.

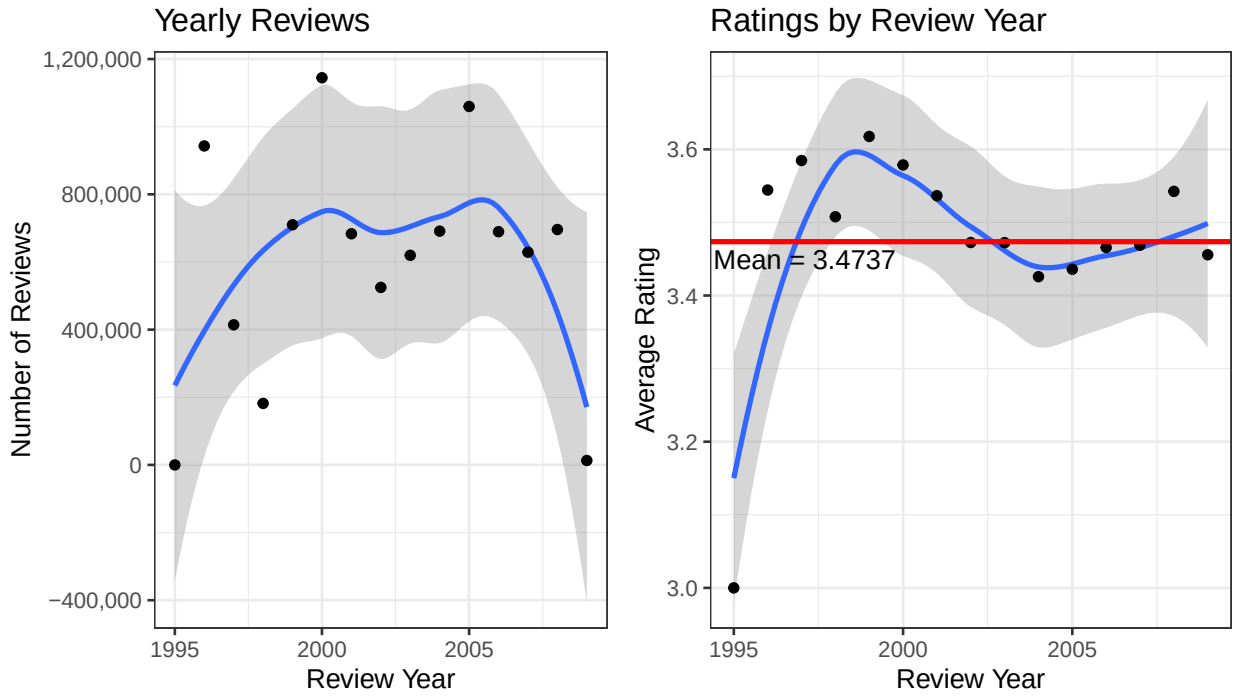
Year of Release

The year of a movie's release is extracted from the movie title. Movie releases cover a 94-year period that ranges from 1915 through 2008. Within those years, a minimum number of reviews is 33 from movies released in 1917 and a maximum number of reviews is 787,116 for movies released in 1995. Likewise, the minimum mean rating is 3.25 from movies released in 1915 and the maximum mean rating is 4.05 from movies released in 1946. The mean rating by year of release is 3.7211, which is higher than the overall average mean of 3.5125. Overall variability in this data is shown in the below figures:



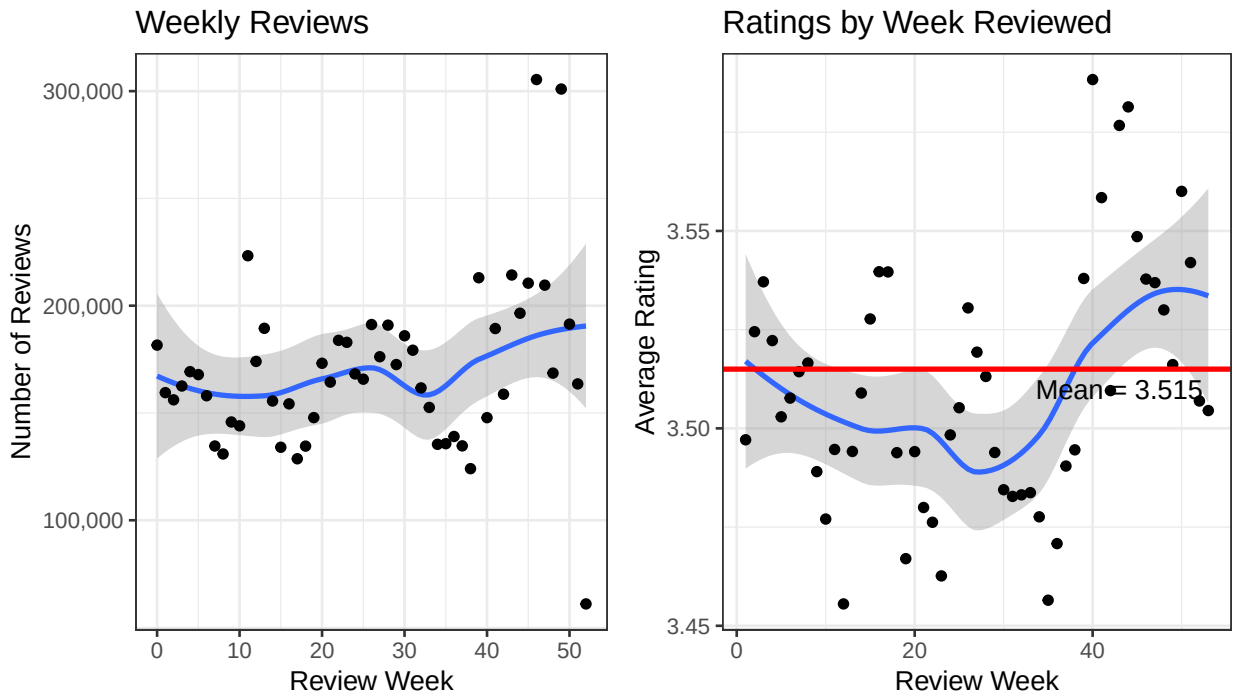
Year of Review

The year that a movie is reviewed and rated is extracted from the timestamp parameter. In the Movielens data, these reviews cover a 15 year period from 1995 through 2009. From the data, variability is again noted with the minimum reviews submitted in 1995 with 2 reviews submitted and the maximum number of reviews occurring in the year 2000 with 114,4666 reviews. Likewise, the minimum mean rating occurred in 1995 with rating of 3.0 and the maximum mean rating occurred in 1999 with a rating of 3.62. The overall mean rating by year of review of 3.4737, which is lower than the overall mean average of 3.5125. Overall variability in this data is shown in the below figures:



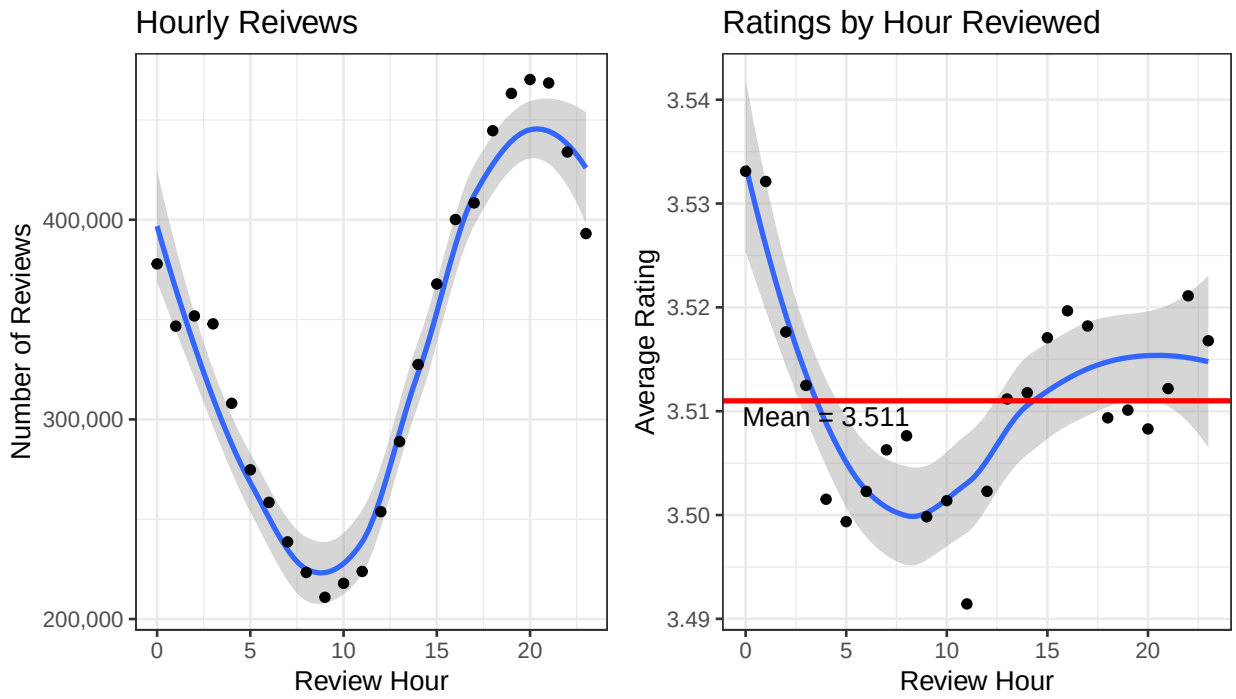
Week of Review

Also extracted from the timestamp is that of the week of the year that the review was submitted. Both reviews and rating scores fluctuate throughout the year and seem to be cyclic in nature. The minimum number of movie reviews were submitted in week 52 (i.e. the week of 21 December) with 54,829 reviews and the maximum number of reviews were submitted at week 46 (i.e. the week of 09 November) with 274,873 reviews. Likewise, the minimum mean rating occurred in week 11 (i.e. the week of 15 March) with a rating score of 3.46 and the maximum mean rating occurred in week 39 (i.e. the week of 21 September) with a rating of 3.59. The overall weekly mean rating was 3.515, which is slightly higher than the overall mean average of 3.5125. Overall variability in this data is shown in the below figures:



Hour of Review

The final value that was used to create a model is that of the hour of the day that reviews were submitted. This can also be extracted from the timestamp; and like the weekly ratings seem to be cyclic in nature. The minimum number of reviews were submitted at 09:00 with 210,888 reviews and the maximum number of reviews submitted at 20:00 with 470,225 reviews. Likewise, the minimum mean rating occurred at 11:00 with a rating of 3.49 and the maximum rating occurred at 00:00 (i.e. midnight) with a rating of 3.53. The hourly mean rating of 3.511 is lower than the overall mean rating of 3.5125. Overall variability in this data is shown in the below figures:



Based on these observations, a model is constructed, expanding on Model 4 and introducing the variable b_t. This new variable accounts for various time effects with λ added to allow for regularization and reduce the impact on the RMSE value from years with few reviews. This equation is written as:

$$Y_{i,u,g,t} = \mu + \lambda_{i,u,g,t} b_i + \lambda_{i,u,g,t} b_u + \lambda_{i,u,g,t} b_g + \lambda_{i,u,g,t} b_t + \epsilon_{i,u,g,t}$$

This model returns the following outputs when it is ran against the test data with the first result depicting the model without regularization (i.e. $\lambda = 0$) and the second result depicting the model results with regularization (i.e. λ set to that of its best tuned value).

```
##    Models                    Predictors Lambda    RMSE
## 8 Model 5     mu+Movie+User+Genre+Time   0.00 0.86397
## 9 Model 5 mu+Movie+User+Genre+Time+Reg   5.25 0.86339
```

Comparing Model 5's unregularized result with that of Model 4 shows a further RMSE reduction of 0.00029. Additionally, Model 5's regularized result reduces the outlier effects by an additional 0.00058.

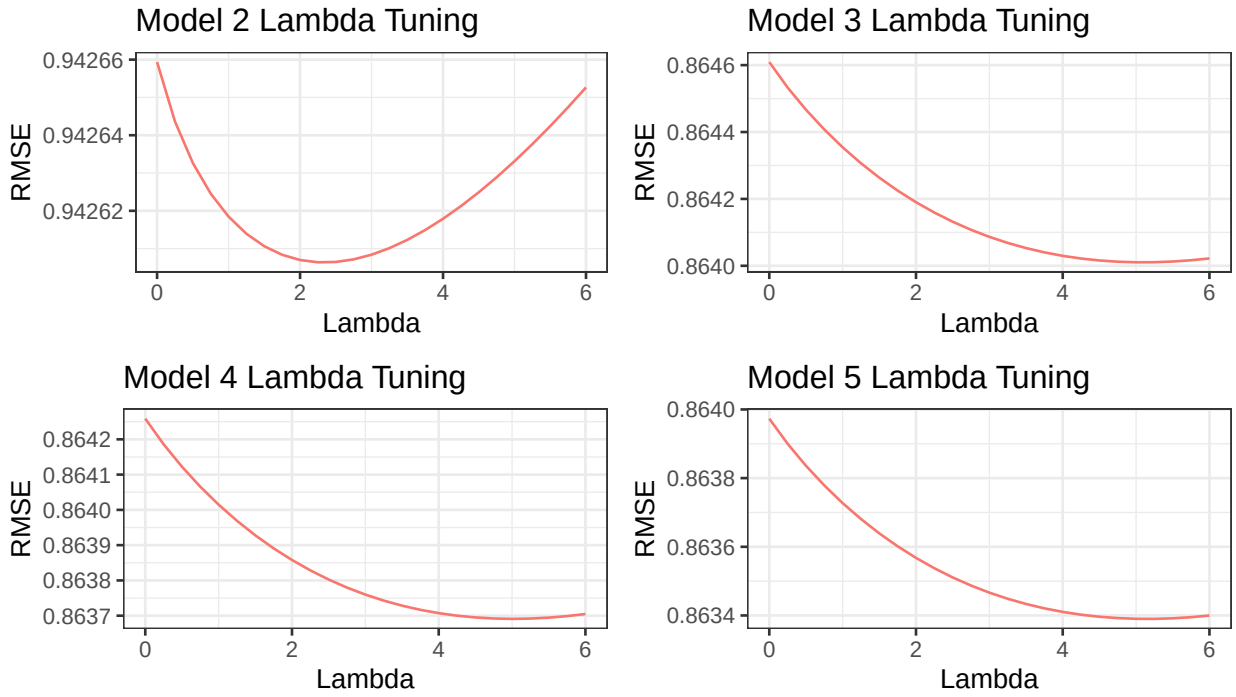
RESULTS

As indicated previously in this report, using an independent data set allows evaluations of the various models' true performance, as training and validating the model from the same data results in over-fitting the model. Utilizing a separate validation data set results in higher RMSE values than that of the results discussed previously from the training and tuning of the models. These results have been separated into non-regularized (i.e. $\lambda = 0$) and regularized (λ is set to its best tuned value) models for ease of comparison.

Models that did not use regularization failed to meet the project's objectives of reducing the RMSE to below 0.86490.

```
##    Models                Predictors Lambda    RMSE
## 1 Model 1                        mu     NA 1.06065
## 2 Model 2                 mu+Movie      0 0.94377
## 4 Model 3            mu+Movie+User      0 0.86610
## 6 Model 4      mu+Movie+User+Genre      0 0.86576
## 8 Model 5 mu+Movie+User+Genre+Time      0 0.86549
```

However, once regularization is introduced with λ tuned to their optimized values, each model showed improvement as the effects of the outliers are reduced.




```
##    Models                    Predictors Lambda    RMSE
## 1 Model 1                            mu     NA 1.06065
## 3 Model 2                 mu+Movie+Reg   2.25 0.94371
## 5 Model 3            mu+Movie+User+Reg   5.00 0.86545
## 7 Model 4      mu+Movie+User+Genre+Reg   5.00 0.86514
## 9 Model 5 mu+Movie+User+Genre+Time+Reg   5.25 0.86485
```

As the results indicate, the regularization result from Model 5 is the only model that met the project's objectives of achieving an RMSE value below that of 0.86490.

```
##    Models                    Predictors Lambda    RMSE
## 1 Model 5 mu+Movie+User+Genre+Time+Reg   5.25 0.86485
```

CONCLUSION

As seen in this project report, model accuracy is impacted by the proper selection of relevant and independent predictors as well as the presence of outliers. While outliers may be addressed through regularization, these models must first be tuned prior to validation; and a large λ will result in overfitting of the models to the test data and selecting independent predictors is important.

Additional studies are warranted as reductions in RMSE may be possible from analysis on movies with multiple genres as well as movies that are prequels or sequals. Cross validation may also assist in reducing the RMSE value, as well as dimension reduction so that standard functions found within R (e.g. linear regression, logistics regression, loess regression, and random forest) may be utilized.

While the project focused on a data set from the movie industry, techniques used may easily be applied to other areas of focus as well, such as product ratings and restaurant ratings, which is used widely by industry to make financial decisions.