HarvardX PH125.9x

MovieLens Project Submission

Chelsea Gorius

Contents

1. Overview

1.1. Introduction

This project aims to utilize the data analysis and machine learning skills taught throughout the HarvardX Data Science Certificate courses. Data analysis and manipulation methods are used to organize and gain a better understanding of the data. Various models were created in order to predict the data by computing effect variables. In addition to this a Regularized Model was used on one of the effect models to observe a possible improvement in the model's Root Mean Squared Error value against it's predictions with an optimized lambda value.

1.2. Project Description

The goal of this project is to develop a movie recommendation system for individual users based on previous movie ratings. The recommendation model should utilize the variable effects present in the data in order to improve its ability to predict user preferences towards other films. Significant data exploration provides an improved general understanding of patterns within the data prior to the effects model creation. Consideration of the effects of each variable helped determined those that would be used to create the recommendation model. Understanding that, though recorded, the 'timestamp' variable will not be as valuable as the 'genres' variable since users often favor specific genres and not timestamp value is important when building the model in order to eliminate unnecessary factors in the final model. It is also valuable to understand that though 'title and 'movieId' are separate variables they can technically be thought of as the same variable. Each movie has one title and one movieId, therefore using both variables in the recommendation model would be pointless. The models developed are being assessed by their root mean squared errors against the validations or test sets of data. The root mean squared error is the standard deviation of prediction errors between the model and the validation or test sets actual values. This project aims to produce a RMSE of less than or equal to 0.8775 when using the final model against the validation set of data.

1.3. Dataset

The data set used to determine the optimal recommendation model, the "edx set" was ninety percent of a whole data set. The other ten percent of data was saved as a "validation set" to be used with the optimal model determined. The edx data set contains 8000071 total observations of 6 variables. This means that there are 8000071 total recorded movie ratings in the edx set. The validation set therefore contains about just under nine hundred thousand recorded movie ratings.

```
> str(edx)
'data.frame':   8000071 obs. of  6 variables:
 $ userId   : int  1 1 1 1 1 1 1 1 1 1 ...
 $ movieId  : num  231 316 355 356 364 377 420 466 586 588 ...
 $ rating   : num  5 5 5 5 5 5 5 5 5 5 ...
 $ timestamp: int  838983392 838983392 838984474 838983653 838983707 838983834 838983834 838984679 838984068 838983339 ...
 $ title    : chr  "Dumb & Dumber (1994)" "Stargate (1994)" "Flintstones, The (1994)" "Forrest Gump (1994)" ...
 $ genres   : chr  "Comedy" "Action|Adventure|Sci-Fi" "Children|Comedy|Fantasy" "Comedy|Drama|Romance|War" ...
>
```

When observing the summary of the edx data set below it should be noted that the userId and movieId distributions values are not actually that valuable. Since each serves as identifiers and therefore without numerical value effect. The rating distribution variables are important however. It should be noted that the rating mean is a 3.5, an entire 1.0 higher than the actual average between the possible 0 and 5 rating range. The title and genres variable length does not refer to the number of unique occurrences in those variables, but simply number of occurrences. That also tells us those variables do not have missing values within the data set.

```
> summary(edx)
     userId        movieId          rating        timestamp           title              genres
 Min.   :    1   Min.   :    1   Min.   :0.500   Min.   :7.897e+08   Length:8000071     Length:8000071
 1st Qu.:18127   1st Qu.:  648   1st Qu.:3.000   1st Qu.:9.468e+08   Class :character   Class :character
 Median :35757   Median : 1834   Median :4.000   Median :1.035e+09   Mode  :character   Mode  :character
 Mean   :35875   Mean   : 4123   Mean   :3.513   Mean   :1.033e+09
 3rd Qu.:53617   3rd Qu.: 3626   3rd Qu.:4.000   3rd Qu.:1.127e+09
 Max.   :71567   Max.   :65133   Max.   :5.000   Max.   :1.231e+09
>
```

The figure on the right first shows that there are a total of 69878 individual users and 10677 unique movies within the edx data set. Below that demonstrates the number of occurrences of each rating value out of the total 8000071 recorded ratings. It can be seen that the top 4 rating values are all three or greater, which accounts for the higher value observed for the rating mean.

```
  n_users n_movies
1   69878    10677
> # Count of the star rat
> edx %>% count(rating) %
# A tibble: 10 x 2
   rating       n
    <dbl>   <int>
 1    4   2301303
 2    3   1885510
 3    5   1235743
 4    3.5  703197
 5    2    632317
 6    4.5  468129
 7    1    307286
 8    2.5  296094
 9    1.5   94674
10    0.5   75818
```

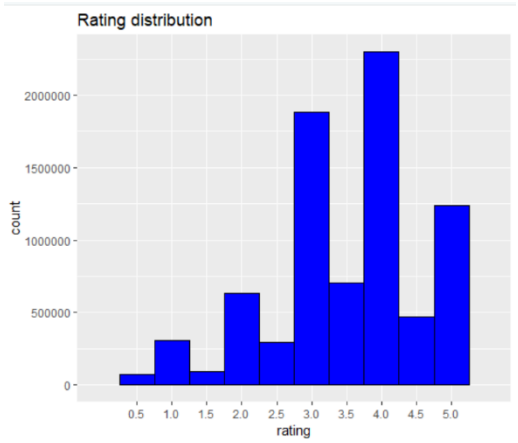
2. Methods and Analysis

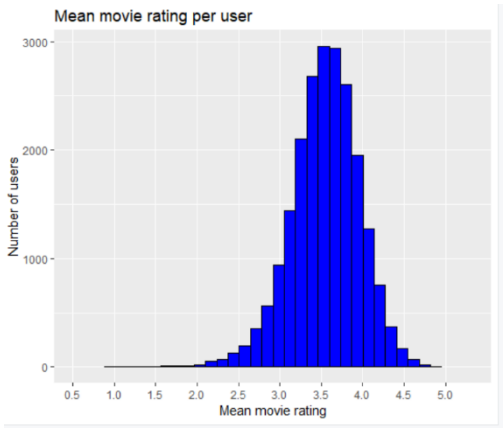
2.1. Data Analysis and Manipulation

As described earlier the entire data source was divided into the edx data set with 90 percent being 8000071 rating observations and the validation data set with ten percent being about just under nine hundred thousand rating observations. The project aims to determine the best model for use against the validation set to produce the best or lowest RMSE value. Whether or not the model is capable of this will be observed by first determine the best model by separating the edx data set into its own test and train data sets. This allows the model to be compared against a training set twice. The models are tested within the edx data set in order to

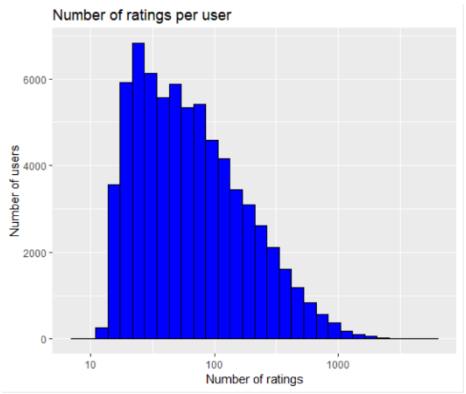
determine the optimal one before being applied to the validation set to observe the actual model success.

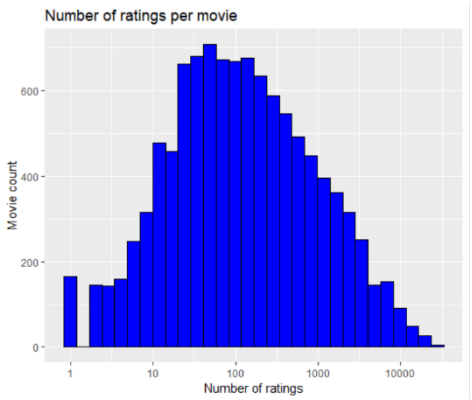
Below the distribution of the rating values, which was observed earlier, can be seen in order to visually interpret the data better. It is clear the observations are skewed to the right towards the higher rating values. From there it can be seen how that distribution of rating values effects the average rating of all the movies. The graph below on the right demonstrates the average movie rating is shaped as a normal distribution with a 3.5 average rating value.





The graphs below observe the distributions of the number of ratings per user as well as per movie. The ratings per user graph shows that most users only rate less than a hundred movies. With the highest quantities below fifty, this means the recommendation system will have to work with fewer observations in most user cases in order to predict user rating values. Lesser observations means the model has less to go off of and often results in larger variability. Though the graph on the right could appear somewhat normal, the increments on the x-axis should be noted. In reality most movies receive fifty to five hundred ratings. Some films boast over ten thousand ratings. This again means the model often has less to work with when determining the effect of specific movies on the recommendation model.

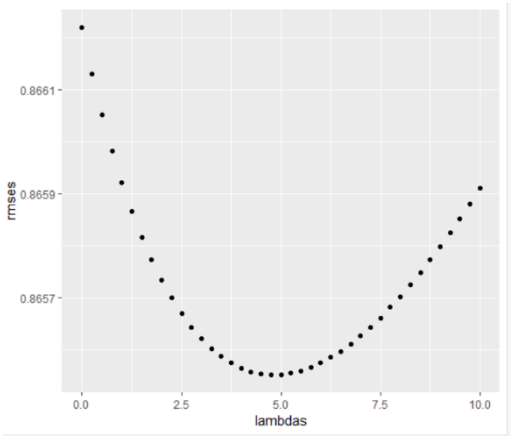




2.2. Modeling Approaches

The recommendation system model aimed to use various variable effects to help predict a users rating of a film. The model begins be first using the average movie rating value as the prediction model as a base rating estimate. From there the model determines the effect each movie could have on its predicted rating. This is done by calculating the individual movieId average ratings and finding it's difference with the overall movie rating average in order to determine the movie effect value. This process is then performed again but with the userId. User's often demonstrate rating patterns, the user effect value is determined by grouping average ratings by userId and subtracting the overall average movie rating as well as that films specific movie effect rating. This process is performed one final time to improve the model with factors via the genres variable. The data is grouped by genre categories and the average ratings are determined, from there the overall film rating average as well as the observation's specific user and movie effects are subtracted in order to determine the genre effect value. In each case the model saw an improvement or a decrease in the model's RMSE value between the edx training set and the edx test set with the addition or a variable effect.

One common factor of error when developing prediction models is overfitting. In order to avoid this regularization was performed on the Movie + User Effect Model. In order to prevent the model from attempting too hard to acknowledge all of the data and focus on the major patterns Regularization applies a lambda penalty to each effect factor. This process is repeated numerous times with a sequence of lambdas in order to determine the optimal lambda which produces the lowest RMSE value. The graph below demonstrates the RMSE versus lambda graph. The graph is concave up parabola with a vertex RMSE value of 5. Meaning the lambda value that produces the lowest RMSE value is 5.



3. Results

The table below demonstrates the RMSE value of each model developed versus either the edx test data set or the validation data set. For the first four observations it is obvious to see how the addition of a variable effect on the model to the rating average improved the RMSE value consistently. One interesting observation is that the regularized movie and user effect model produced a lower RMSE than the model with movie, user, and genre effects. Since that model with three variable effects has a lower RMSE than the one with two we can conclude that the third variable did improve the model and not worsen it overall with overfitting. The results of the regularized model with variables suggest that the regularized three variable model would produce an even lower RMSE value against the edx test, a likely next step.

method	RMSE
Average	1.0600537
Movie Effect Model	0.9429615
Movie + User Effects Model	0.8646844
Movie + User + Genre Effects Model	0.8643242
Regularized Movie + User Effect Model - test dataset	0.8641362
Regularized Movie + User Effect Model - Validation dataset	0.8648177

In the end it was the regularized movie and user effect model that produced the lowest RMSE value against the edx test data set. That model was then applied to the validation set, the original ten percent of the data available at the start of the project. When applied here the model produced an RMSE of 0.8648. This value though higher than the one produced against the edx test set met the project goal of being less than or equal to 0.8775

4. Conclusion

In conclusion the model accomplished the set goal of obtaining a RMSE value less than or equal to 0.8775 with a value of 0.86848. The visualization of the data provided increased understanding of likely variable effects as well obstacles that ay have been faced in analysis and model training. The different models produced demonstrated the valuable effect of increased number variable effects (within the zero to three range). In addition to that the project demonstrates how overfitting can affect a model's success and that regularization can force a model with less variables to be more accurate. This case specifically points to the model which produced the lowest RMSE value against the edx test data set and met the RMSE goal against the validation test set, the regularized movie and user effect model.