Movielens_Project_Report

Curios_i

04/05/2021

Contents

1 Introduction

Though recommender systems are not new, the interest in a movie rating recommender system was fueled by Netflix challenge announced in 2006 and finally won by BellKor's Pragmatic Chaos team in 2009. The goal of this project is to create a movie recommendation system using the movielens dataset. The movielens dataset comprises of 10M movie ratings from Netfflix data as compared to more than 100M used in Netflix challenge. To build a recommender system, this report will discuss two solutions. Please note that this is a learning project and not a research project. Although the Solution#2 is more efficient and robust, it is kind of plug and play, that's why Solution#1 is also included which dissects about the main underlying theory of a recommender system. In the end, we shall compare both solutions and their limitations. Though both soutions have achieved the target RMSE, we shall still emphasize that the intent of this project and report is learning and sharing, not winning any competition. So, engjoy...

The project report is divided into following sections:

- Data Analysis : to analyze the Movielens data and the nature of the rating prediectio problem.
- Method : to discuss the method employed to construct a recommendation system based on the data analysis
- Results: to present modelling results and discuss the model performance.
- Conclusion: to give brief summary of the report, its limitations and future work.

2 Data Analysis

2.1 Grabbing the data

The 10M movie rating data is downloaded from this link. Here is the code to convert this data into a data frame and then divide into a edx and validation data set.

```
if(!require(tidyverse)) install.packages("tidyverse")
if(!require(caret)) install.packages("caret")
if(!require(data.table)) install.packages("data.table")
if(!require(data.table)) install.packages("Matrix")
if(!require(data.table)) install.packages("recommenderlab")

#tcrossprod() is used from library Matrix, while
#funkSVD() is used from library recommenderlab

library(tidyverse)
library(caret)
library(data.table)
library(Matrix)
library(recommenderlab)
library(knitr)

# MovieLens 10M dataset:
# https://grouplens.org/datasets/movielens/10m/
# http://files.grouplens.org/datasets/movielens/ml-10m.zip

dl <- tempfile()
download.file("http://files.grouplens.org/datasets/movielens/ml-10m.zip", dl)

ratings <- fread(text = gsub("::", "\t",readLines(unzip(dl,"ml-10M100K/ratings.dat"))),
```

```
col.names = c("userId", "movieId", "rating", "timestamp"))
movies <- str_split_fixed(readLines(unzip(dl, "ml-10M100K/movies.dat")), "\\::", 3)
colnames(movies) <- c("movieId", "title", "genres")

movies <- as.data.frame(movies) %>% mutate(movieId = as.numeric(movieId),
title = as.character(title), genres = as.character(genres))

movielens <- left_join(ratings, movies, by = "movieId")

# Validation set will be 10% of MovieLens data
set.seed(1, sample.kind="Rounding") # if using R 3.5 or earlier, use `set.seed(1)`
test_index <- createDataPartition(y = movielens$rating, times = 1, p = 0.1, list = FALSE)
edx <- movielens[-test_index,]
temp <- movielens[test_index,]

# Make sure userId and movieId in validation set are also in edx set
validation <- temp %>%
  semi_join(edx, by = "movieId") %>%
  semi_join(edx, by = "userId")

# Add rows removed from validation set back into edx set
removed <- anti_join(temp, validation)
edx <- rbind(edx, removed)

rm(dl, ratings, movies, test_index, temp, movielens, removed)
```

Next, we divide edx data frame into edx_train and edx_test using the following code so that we can tune and test our module without touching the validation data frame. But, before doing that, we increase the virtual memory to 200,000MB so that it can handle this huge amount of data.

```
memory.limit(size=200000) #sets virtual memory size to 200,000Mb to process large data
set.seed(1, sample.kind="Rounding")
#Partition edx into 80% edx_train and 20% edx_test so
#that we can tune the model without touching the validation data
test_index <- createDataPartition(y = edx$rating, times = 1, p = 0.2, list = FALSE)
edx_train<-edx[-test_index,]
temp<-edx[test_index,]

#make sure that the movieId and userId in edx_train are also in edx_test
edx_test<-temp%>%semi_join(edx_train,by="movieId")%>%semi_join(edx_train,by="userId")
#Add removed records from edx_test back to edx_train
removed<-anti_join(temp,edx_test)
edx_train<-rbind(edx_train,removed)
#Remove temporary variables to free up the memorylib
rm(removed,temp,test_index)
```

2.2 Analyzing the Data and Deciding the Model

The next step is to analye the data and decide which model to use. However, before doing that, we need to decide our loss function. We shall use root-mean-squre-error as our loss function. So, let's define our loss function in the following code:

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

There are two main challenges with data in this project

- The size of data
- The nature of the model

The first temptation is to use a machine learning model on this data, like linear regression, logistic regression, LDA, QDA etc. The problem with this approach is that userId and movieId are just arbitrary variables. In fact, userId was used by Netflix to protect the privac of users. There is no cause and effect relationship between userId vs rating or movieId vs rating. In fact, if I take userId from 1 to 1000 and 1001 to 2000 and swap them, it should not affect the prediction model at all. To prove this point, let's see the correlation.

```
cor(edx$userId,edx$rating)
```

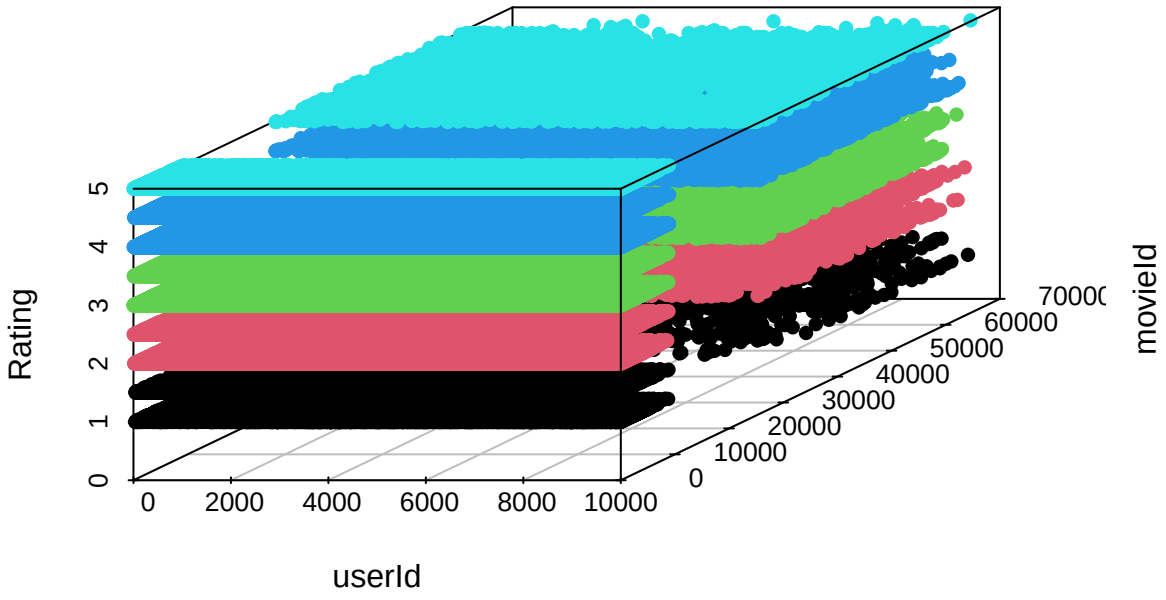
```
## [1] 0.002313643
```

```
cor(edx$movieId,edx$rating)
```

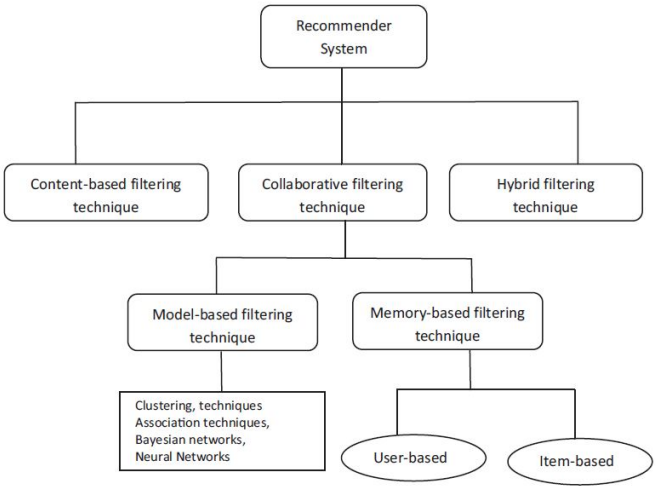
```
## [1] -0.006535696
```

Another issue is the size of the data. For such a large dataset if you try to use any model using normal r package (e.g. caret), it will crash your computer, or you need a very large computing resource, which is out of reach of most of the students. One solution to this problem is to use stochastic gradient descent algorithm, which we shall discuss later in this report. In fact, we created a liner regression model of this dataset using "SGD" package, but the RMSE obtained was greater than 1.5, for the reasons explained above, so we are not even including that in this report. Now, to have a visual sense of dataset, let's first dissect the edx_train into a smaller dataset and then make a 3D plot.

```
train_small<-edx_train%>%filter(userId<10000)
if(!require(scatterplot3d)) install.packages("scatterplot3d")
library(scatterplot3d)
scatterplot3d(x=train_small$userId,z=train_small$rating,y=train_small$movieId,
color = train_small$rating,xlab="userId",ylab="movieId",zlab="Rating",pch=16)
```



It is clear from the plot that applying a normal machine learning technique as described above will not be very successful on this data. ## Recommender System Modeling Techniques Following chart shows well known recommender system modeling techniques.



2.2.1 Conent Based Filtering

Content based technique emphasizes more on the analysis of the attributes of items in order to generate predictions. When items, like web pages, publications and news are to be recommended, content-based filtering technique is the most successful. Content based filtering requires a lot of meta data about items.

Since in the case of movielens data the only meta data somewhat useful is the movie genera, we are not considering content-based filtering in this project.

[3] chapter 3 discusses an example of movie reviews either positive or negative based on 50,000 reviews from IMDB data set using deep learning. The same model can be extended to predict movie ratings from 1-5 based on reviews. As mentioned above, we don't have enough data to apply this filtering in this project.

2.2.2 Collborative Filtering

Collaborative filtering is a prediction technique for content that cannot easily and adequately be described by metadata, such as movies and music. Collaborative filtering technique works by building a database (user-item matrix) of preferences for items by users. It then matches users with relevant interest and preferences by calculating similarities between their profiles to make recommendations. Such users build a group called neighborhood. An user gets recommendations to those items that he has not rated before but that were already positively rated by users in his neighborhood. The technique of collaborative filtering can be divided into two categories:memory-based and model-based.

2.2.2.1 Memory Based Collaborative Filtering Memory-based CF can be achieved in two ways through user-based and item-based techniques. User based collaborative filtering technique calculates similarity between users by comparing their ratings on the same item and it then computes the predicted rating for an item by the active user as a weighted average of the ratings of the item by users similar to the active user where weights are the similarities of these users with the target item. Item-based filtering techniques compute predictions using the similarity between items and not the similarity between users.

In our Solution#1 we have used memory based collaborating filtering along with regularization on top of overall average rating.

2.2.2.2 Model Based Techniques Model based techniqes employ previous ratings to learn a model in order to improve the performance of collaborative filtering. These techniques can quickly recommend a set of items for the fact that they use pre-computed model. The most popular model based techniqe in recommended system is Singular Value Decomposition (SVD). In both our Solution#1 and Solution#2 matrix factorization by SVD is used as a model based CF filtering.

3 Solution#1

Solution#1 is based on [1] and [2]. The model for predicted ratings have four components, calculated from edx_train

$$Y_{u,i} = \mu + b_i + b_u + r_{u,i} + \xi_{u,i}$$

*1. μ - mean of ratings

*2. b_i - average rating of a movie i, with L2 regularization

*3. b_u - average rating given by a user I, with L2 regularization

*4. $r_{u,i}$ - residual ratings calculated from above. we can write

$$pred = r_{u,i} + \xi_{u,i} = Y_{u,i} - \mu + b_i + b_u$$

In our computation we shall use variable `pred` to represent the left hand side of the equation and will use matrix factorization to estimate that. `pred` represents the fact that group of users have similar rating

patterns for group of movies, i.e. it is the interaction between users and movies. We shall estimate it by using singular value decomposition.L

The first step is to calculate the first component, i.e. μ - mean of ratings

```
mu<-mean(edx_train$rating)
```

Next step is to calculate values of b_i and b_u using L2 regularization. We shall use equations given in [1] and [2]

$$b_i(\lambda) = \frac{1}{\lambda + n_i} \sum_{a=1}^{n_i} (Y_a - \mu)$$

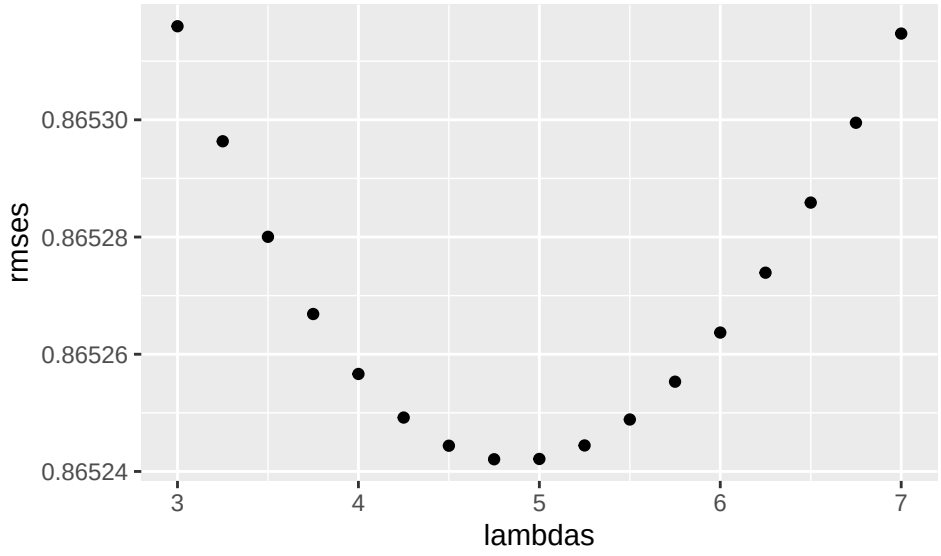
$$b_u(\lambda) = \frac{1}{\lambda + n_u} \sum_{a=1}^{n_u} (Y_a - \mu - b_i)$$

Since b_i and b_u are function of λ, we first define an range of lambdas and then find rmses for every value in the lambdas vector.

```
lambdas <- seq(3, 7, 0.25)
# calculates rmses for the above values of lambdas to tune the model
rmses<-sapply(lambdas,function(l){
  b_i<-edx_train%>%group_by(movieId)%>%summarise(b_i=sum(rating-mu)/(n()+l))
  b_u<-edx_train%>%left_join(b_i,by="movieId")%>%
    group_by(userId)%>%
    summarise(b_u=sum(rating-b_i-mu)/(n()+l))
  predicted_ratings_test<-edx_test%>%left_join(b_i,by="movieId")%>%
    left_join(b_u,by="userId")%>%
    mutate(pred=mu+b_i+b_u)%>%
    pull(pred)
  return(RMSE(predicted_ratings_test,edx_test$rating))})
```

Now, plot of rmses vs lambdas

```
library(ggplot2)
qplot(lambdas,rmses)
```



We find the the minimum rmse at this point is

```
min(rmses)
```

```
## [1] 0.8652421
```

at the value of lambda

```
lambda<-lambdas[which.min(rmses)]
lambda
```

```
## [1] 4.75
```

Now we shall calculate b_i and b_u for edx_train. Those b_i and b_u are then used to calculate predicted ratings for edx_train and edx_test. Predicted ratings for edx_test is later used along with funk svd predictions to tune the model.

Predicted ratings for edx_train are used to calculate the residual ratings. Funk svd matrix factorization is applied on those residual edx_train ratings to build the model.

Here is the code:

```
# calculates b_i and b_u for edx_train. Since all users and movies in edx_test and
#validation are present in edx_train, these values of b_i and b_u are used later  to
#calculate the predicted ratings for those datasets
b_i<-edx_train%>%group_by(movieId)%>%
  summarise(b_i=sum(rating-mu)/(n()+lambda))
b_u<-edx_train%>%left_join(b_i,by="movieId")%>%
  group_by(userId)%>%
  summarise(b_u=sum(rating-b_i-mu)/(n()+lambda))
# calculates predicted ratings for the edx_test based on b_i and b_u
predicted_ratings_test<-edx_test%>%
  left_join(b_i,by="movieId")%>%
  left_join(b_u,by="userId")%>%
  mutate(pred=mu+b_i+b_u)%>%
```

```
  pull(pred)
#calculates predicted ratings for edx_train, so that we can calculate
#residuals for further matrix factorization
predicted_ratings_train<-edx_train%>%
  left_join(b_i,by="movieId")%>%
  left_join(b_u,by="userId")%>%
  mutate(pred=mu+b_i+b_u)%>%
  pull(pred)
#calculated residual ratings after subtracting the predicted ratings
#from edx_train ratings
edx_train<-edx_train%>%
  mutate(predicted_ratings_train=predicted_ratings_train,
         resid=rating-predicted_ratings_train)
```

Now, we shall convert `resid` into a matrix, where each row will represent a user and each column will represent a movie.

```
y<-edx_train %>% select(userId,movieId,resid) %>% spread(movieId,resid) %>% as.matrix()
rownames(y)<- y[,1]
y<-y[,-1]
```

3.1 Funck SVD

A really smart realization made by the guys who entered the Netflix's competition (notably Simon Funk) was that the users' ratings weren't just random guesses. Raters probably follow some logic where they weight the things they like in a movie (a specific actress or a genre) against things they don't like (long duration or bad jokes) and then come up with a score.

That process can be represented by a linear formula of the following kind:

$$R_{u,i} = UV^T$$

Here,

U is a matrix of rows equal to number of users (or number of rows of $r_{u,i}$) and k columns

V is a matrix of rows equal to number of items/movies (or number of columns of $r_{u,i}$) and k columns

k is number of features we want to extract

$R_{u,i}$ is a user rating matrix, with u users and i ratings.

In a recommender system database, every user doesn't rate every movie, so $R_{u,i}$ is a sparse matrix, where most of the elements are unknown.

U and V can be found in such a way that the square error difference between their cross product and known rating in the user-item matrix is minimum.

If we consider u as a row vector of U matrix and v as a row vector of V matrix, we can write

$$\text{expected rating} = \hat{r_{u,i}} = uv^T$$

Our goal is to minimize the following for each known rating.

$$min(u,v)\sum(r_{u,i} - uv^T)$$

For our model to be able to generalize well and not over-fit the training set, we introduce a penalty term to our minimization equation. This is represented by a regularization factor γ multiplied by the square sum of the magnitudes of user and item vectors, in case of L2 regularization (ridge regression)

$$min(u,v)\sum(r_{u,i} - uv^T) + \gamma(||u||^2 + ||v||^2)$$

While in case of L1 regularization,

$$min(u,v)\sum(r_{u,i} - uv^T) + \gamma(||u|| + ||v||)$$

funkSVD is a function from `recommenderlab` library. In the funkSVD function, we use L1 regularization, where all elements of U and V matrices are set to zero initially. Following is the usage of the function with default values of arguments.

funkSVD(x, k = 10, gamma = 0.015, lambda = 0.001, min_improvement = 1e-06, min_epochs = 50, max_epochs = 200, verbose = FALSE)

Arguments

Parameter	Description
x	a matrix, potentially containing NAs.
y	number of features (i.e. rank of approximation)
gamma	regularization term
lambda	learning rate
min_improvement	required minimum improvement per iteration
min_epochs	minimum number of iterations per feature
max_epochs	maximum number of iterations per feature
verbos	show progress

3.2 Tuning funkSVD model

Since running funkSVD requires a lot of computational power and memory, we shall only use k = 3 features since that will give us less than the target RMSE. Using more features may definitely improve the RMSE.

Besides that, there a other hyperparameters, like epochs, gamma, lambda and min_improvement. We shall keep the learning rate lambda default, however, for regularization, we shall tune our model for three values of gamma. We define a corresponding vector f_rmse and initialize it to store the RMSE outcomes of three models.

```
# gammas are L1 regularization terms in svd gradient descent algoritm
# we shall calculate rmses for each of these gammas and will determin which one gives optimal tuning
gammas<-c(0.015,0.025,0.035)
# f_rmses stores rmses calculated for respective values of gammas
f_rmses<-c(1,1,1)
```

3.2.1 For gamma = 0.015

As mentioned above, funkSVD function returns us two fsvd$Uandfsvd$V matrices. Cross product of these two matrices gives us our prediction matrix for residuals. We add those predictions to predicted_ratings_test. Remember that

$$predicted\_ratings\_test = \mu + b_i + b_u$$

```
#############################################################################
# runs the Simon Funk's gradient descent algorithm to factorize matrix of residuals
# this will take several hours
fsvd<-funkSVD(y, k=3, gamma=gammas[1], lambda=0.001, verbose=TRUE)
# y_hat is prediction matrix from SVD
y_hat_0.015<-tcrossprod(fsvd$U,fsvd$V)
# Assigns row and column names to prediction matrix so that predicted rating can be pulled from the mat
rownames(y_hat_0.015)<-rownames(y)
colnames(y_hat_0.015)<-colnames(y)
# creates a placeholder for pred = predicted ratings for residuals from SVD
pred<-rep(0,length(edx_test$userId))
# fill vector pred from the prediction matrix for respective userId and movieId
for(i in 1:length(edx_test$userId)){
  pred[i]<-y_hat_0.015[as.character(edx_test$userId[i]),as.character(edx_test$movieId[i])]}
# calculated predicted ratings form edx_test for that particular gamma
predicted_ratings_0.015<-predicted_ratings_test+pred
# calculates rmse for that particular gamma
f_rmses[1]<-RMSE(predicted_ratings_0.015,edx_test$rating)
#[1] 0.8260231
```

The RMSE calculated for gamma=0.015 is

```
f_rmses[1]
```

```
## [1] 0.8260231
```

3.2.2 For gamma = 0.025

As mentioned above, funkSVD function returns us two fsvd$U and fsvd$V matrices. Cross product of these two matrices gives us our prediction matrix for residuals. We add those predictions to predicted_ratings_test. Remember that

$$predicted\_ratings\_test = \mu + b_i + b_u$$

```
#############################################################################
# runs the Simon Funk's gradient descent algorithm to factorize matrix of residuals
fsvd<-funkSVD(y, k=3, gamma=gammas[2], lambda=0.001, verbose=TRUE)
# y_hat is prediction matrix from SVD
y_hat_0.025<-tcrossprod(fsvd$U,fsvd$V)
# Assigns row and column names to prediction matrix so that predicted rating can be pulled from the mat
rownames(y_hat_0.025)<-rownames(y)
colnames(y_hat_0.025)<-colnames(y)
pred<-rep(0,length(edx_test$userId))
# fill vector pred from the prediction matrix for respective userId and movieId
for(i in 1:length(edx_test$userId)){
  pred[i]<-y_hat_0.025[as.character(edx_test$userId[i]),as.character(edx_test$movieId[i])]}
# calculated predicted ratings form edx_test for that particular gamma
predicted_ratings_0.025<-predicted_ratings_test+pred
# calculates rmse for that particular gamma
f_rmses[2]<- RMSE(predicted_ratings_0.025,edx_test$rating)
#0.8257668
```

The RMSE calculated for gamma=0.025 is

```
f_rmses[2]
```

```
## [1] 0.8257668
```

3.2.3 For gamma = 0.035

As mentioned above, funkSVD function returns us two fsvd$U and fsvd$V matrices. Cross product of these two matrices gives us our prediction matrix for residuals. We add those predictions to predicted_ratings_test. Remember that

$$predicted\_ratings\_test = \mu + b_i + b_u$$

```
#############################################################################
# runs the Simon Funk's gradient descent algorithm to factorize matrix of residuals
fsvd<-funkSVD(y, k=3, gamma=gammas[3], lambda=0.001, verbose=TRUE)
y_hat_0.035<-tcrossprod(fsvd$U,fsvd$V)
rownames(y_hat_0.035)<-rownames(y)
colnames(y_hat_0.035)<-colnames(y)
pred<-rep(0,length(edx_test$userId))
for(i in 1:length(edx_test$userId)){
  pred[i]<-y_hat_0.035[as.character(edx_test$userId[i]),as.character(edx_test$movieId[i])]}
predicted_ratings_0.035<-predicted_ratings_test+pred
f_rmses[3]<- RMSE(predicted_ratings_0.035,edx_test$rating)
#[1] 0.8258410
```

The RMSE calculated for gamma=0.035 is

```
f_rmses[3]
```

```
## [1] 0.825841
```

3.2.4 Calculating Final RMSE for Validation

After calculating RMSEs for all three values of gammas, we select the gamma which gave us lowest RMSE.

```
#calculates the optimal gamma from rmses results
gamma_min<-gammas[which.min(f_rmses)]
gamma_min
```

```
## [1] 0.025
```

```
# assings the final predicttion matrix from the min rmse gamma
y_hat_final<-y_hat_0.025
```

Next, we calculate predicted ratings for validation set using μ, b_i and $b_u from edx_train.

```r
# predicted ratings for validation set is first calculated from mu, b_i and b_u from edx_train
predicted_ratings_valid<-
  validation%>%left_join(b_i,by="movieId")%>%
  left_join(b_u,by="userId")%>%
  mutate(pred=mu+b_i+b_u)%>%
  pull(pred)
```

From here, we calculate first RMSE before applying SVD residual calculations

```r
# This finds rmse before applying SVD to residual
RMSE(validation$rating,predicted_ratings_valid)
```

```
## [1] 0.8657012
```

Following code now calculates validation RMSE

```r
# A residual vector placeholder for each user in the validation
pred<-rep(0,length(validation$userId))
# pulls the predicted value of residual from the prediction matrix
for(i in 1:length(validation$userId)){
  pred[i]<-y_hat_final[as.character(validation$userId[i]),as.character(validation$movieId[i])]}
#calculates the final predicted ratings for validation set
predicted_ratings_valid<-predicted_ratings_valid+pred
# calculates final rmse
RMSE(predicted_ratings_valid,validation$rating)
```

```
## [1] 0.8252377
```

4 Bibliography

[1] Yehuda Koren. The BellKor Solution to the Netflix Grand Prize

[2] Rafael A. Irizarry. Introduction to Data Science

[3] François Chollet and Joseph J. Allaire. Deep Learning with R