MovieLens

Simone Trindade Steel

17/05/2020

Executive Summary

This report shows the RMSE evaluation of movie ratings predictions using the "edx" data set for training and the "validation" set for evaluation.

After exploring the relationships between predictors, these are the key insights used in the construction of the prediction model:

1. Ratings for films in recent decades have declined;
2. There no indication that genre influences rating and
3. Categorisation models are not appropriate for the prediction, therefore an adjusted Naive Bayes approach was taken in order to account for movie, user and decade biases in the predictors.

The code provided by EDX to load the approapriate data and libraries has been used here.

```
## Loading required package: tidyverse

## -- Attaching packages ------------------------------------------------------------------------ tidyver

## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  3.0.0     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## Warning: package 'tibble' was built under R version 3.6.2

## -- Conflicts ------------------------------------------------------------------------ tidyverse_con
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Loading required package: caret

## Loading required package: lattice

## Warning: package 'lattice' was built under R version 3.6.2

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift

## Loading required package: data.table

##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

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

```
## Joining, by = c("userId", "movieId", "rating", "timestamp", "title", "genres")
```

Glancing at the edx dataset used for training

Structure and sample data:

```r
str(edx)
```

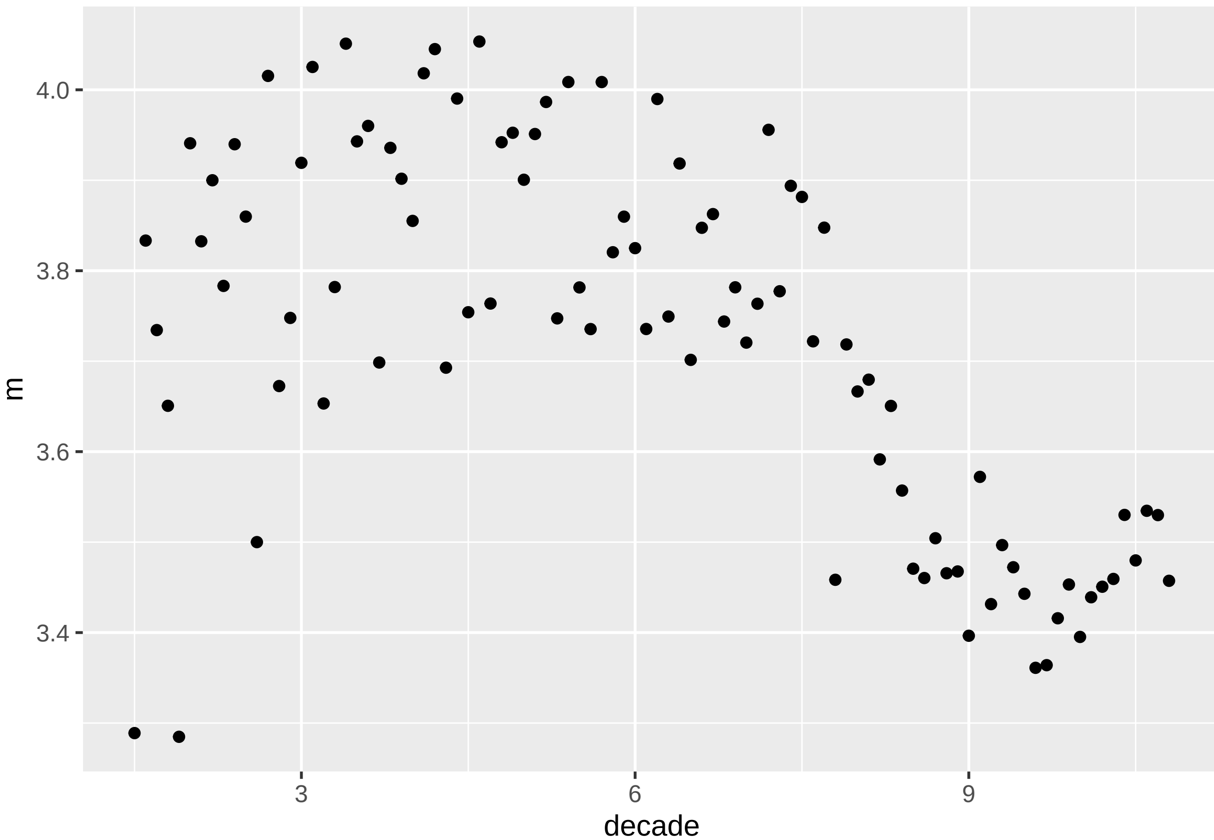
```
## 'data.frame':    9000055 obs. of  6 variables:
##  $ userId   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ movieId  : num  122 185 292 316 329 355 356 362 364 370 ...
##  $ rating   : num  5 5 5 5 5 5 5 5 5 5 ...
##  $ timestamp: int  838985046 838983525 838983421 838983392 838983392 838984474 838983653 838984885 83
##  $ title    : chr  "Boomerang (1992)" "Net, The (1995)" "Outbreak (1995)" "Stargate (1994)" ...
##  $ genres   : chr  "Comedy|Romance" "Action|Crime|Thriller" "Action|Drama|Sci-Fi|Thriller" "Action|Ad
```

```r
head(edx)
```

```
##   userId movieId rating timestamp                         title
## 1      1     122      5 838985046              Boomerang (1992)
## 2      1     185      5 838983525               Net, The (1995)
## 4      1     292      5 838983421               Outbreak (1995)
## 5      1     316      5 838983392               Stargate (1994)
## 6      1     329      5 838983392 Star Trek: Generations (1994)
## 7      1     355      5 838984474       Flintstones, The (1994)
##                           genres
## 1                 Comedy|Romance
## 2          Action|Crime|Thriller
## 4   Action|Drama|Sci-Fi|Thriller
## 5        Action|Adventure|Sci-Fi
## 6 Action|Adventure|Drama|Sci-Fi
## 7        Children|Comedy|Fantasy
```

Analysing rating and time relationship

Ratings and time are weakly, but negatively correlated. Recent films have slightly lower ratings than older ones.



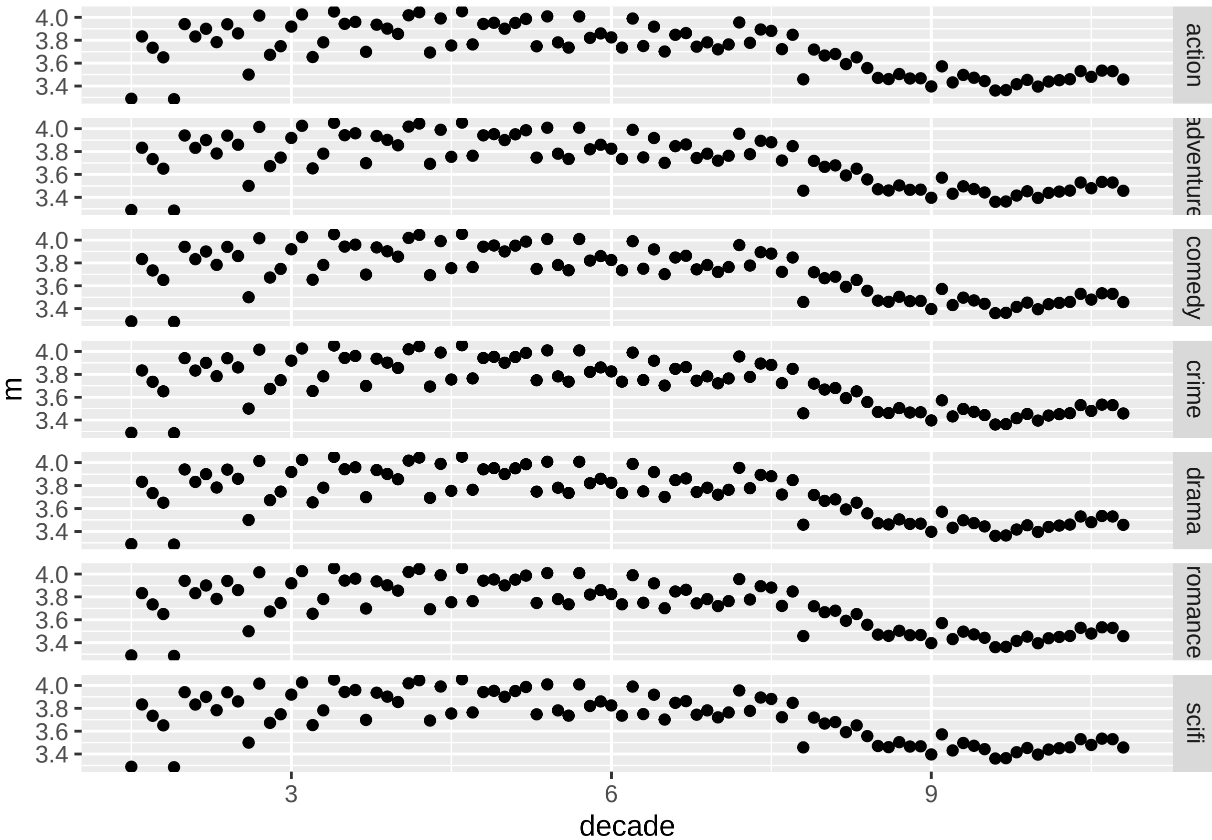
Analysing rating and genre relationship

Looking into the most common genres (the ones with over one million reviews), ratings do not seem to vary across them.

```
## # A tibble: 17 x 2
##    genre_name        s
##    <chr>         <dbl>
##  1 drama       3910127
##  2 comedy      3540930
##  3 action      2560545
##  4 thriller    2325899
##  5 adventure   1908892
##  6 romance     1712100
##  7 scifi       1341183
##  8 crime       1327715
##  9 fantasy      925637
## 10 children     737994
## 11 horror       691485
## 12 mystery      568332
## 13 war          511147
## 14 animation    467168
## 15 musical      433080
## 16 western      189394
## 17 documentary   93066
```

Exploring genre and time together

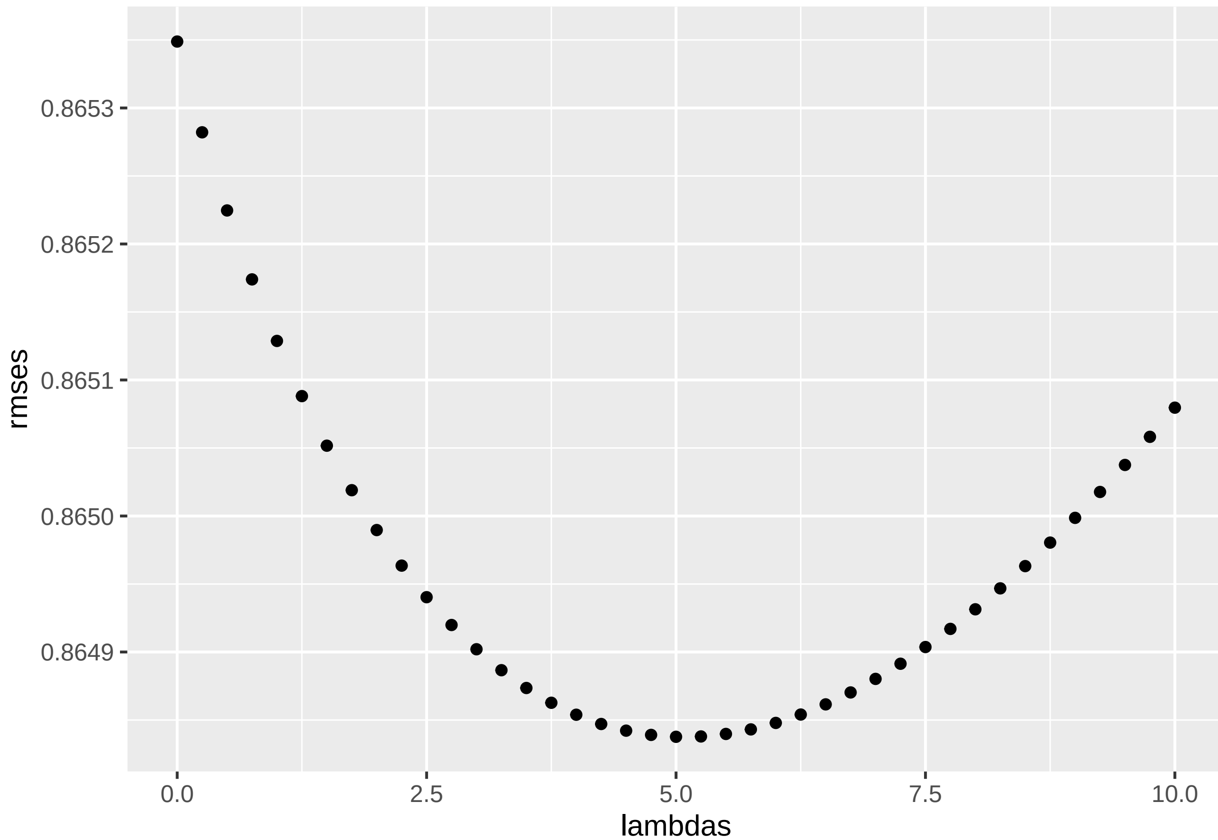
Similar conclusion is drawn when genre by decade analysis is performed. That is, no significant difference in trend.



Chosing an appropriate model: improving on Naive Bayes (movie, user and time biases), with regularisation to control for differences in number of ratings per movie

RMSE Results

Utilising the best regularisation parameter that minimises RMSE.




```
## [1] "Lamba = " "5"
```

method	RMSE
Regularized Movie, User and Time Effect Model	0.8648377