MovieLens

Simone Trindade Steel

25/07/2020

1. Introduction Executive Summary

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

After exploring and visualising 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 is 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 appropriate data and libraries has been used in the generation of this report.

For ease of reference, the section of the Report.R file is commented as "# Create edx set, validation set (final hold-out test set)".

2. Method and analysis

The initial step for analysis was to glance at the edx (training) dataset. Its structure and sample data can be seen here:

```
str(edx)
```

```
## Classes 'data.table' and '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|Ac
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
head(edx)
```

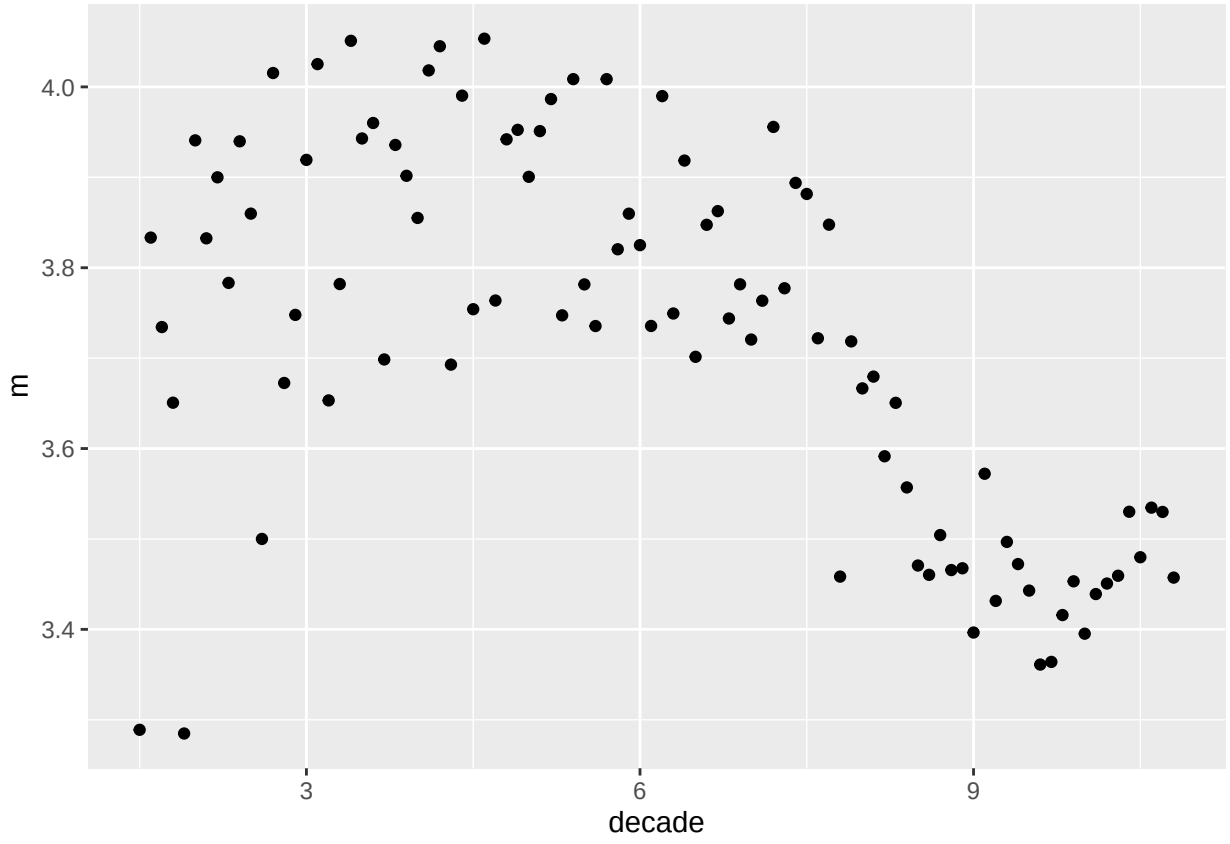
```
##    userId movieId rating timestamp                         title
## 1:      1     122      5 838985046               Boomerang (1992)
## 2:      1     185      5 838983525                Net, The (1995)
## 3:      1     292      5 838983421                Outbreak (1995)
## 4:      1     316      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
```

2.1. Analysing rating and time relationship

The section of the Report.R file that refers to the data analysis is also commented as "# Analysing rating and time relationship".

As seen in the graph below, ratings and time are weakly, but negatively correlated. Recent films have slightly lower ratings than older ones.

```
## `summarise()` ungrouping output (override with `.groups` argument)
```



2.2. Analysing rating and genre relationship

The section of the Report.R file that refers to the data analysis is also commented as "# Analysing rating and genre relationship".

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

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

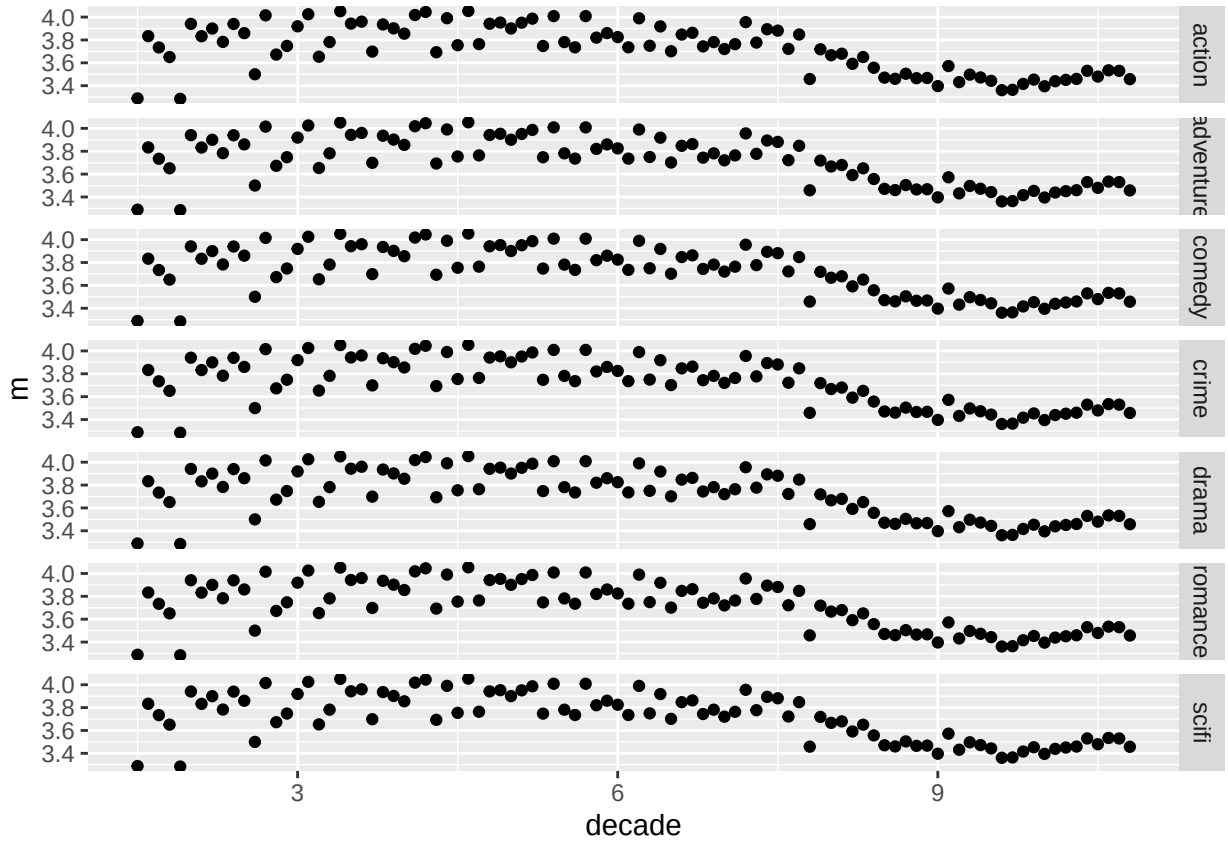
```
## # A tibble: 17 x 2
##    genre_name        s
##    <chr>         <dbl>
##  1 drama       3910127
##  2 comedy      3540930
##  3 action      2560545
##  4 thriller    2325899
##  5 adventure   1908892
##  6 romance     1712100
##  7 scifi       1341183
##  8 crime       1327715
##  9 fantasy      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
```

2.3. Exploring genre and time together

Similar conclusion is drawn when genre by decade analysis is performed. That is, no significant difference in trend, as demonstrated in the graph below. Genre will be discarded as a predictor.

See "# Exploring genre and time together" in the Report.R script.

```
## `summarise()` regrouping output by 'genre_name' (override with `.groups` argument)
```



2.4. Chosing an appropriate predictive model

This report will attempt to improve on Naive Bayes, using movie, user and time as biases. Regularisation to control for differences in number of ratings per movie will also be applied.

The section of the Report.R script is commented as "# Chosing an appropriate predictive model".

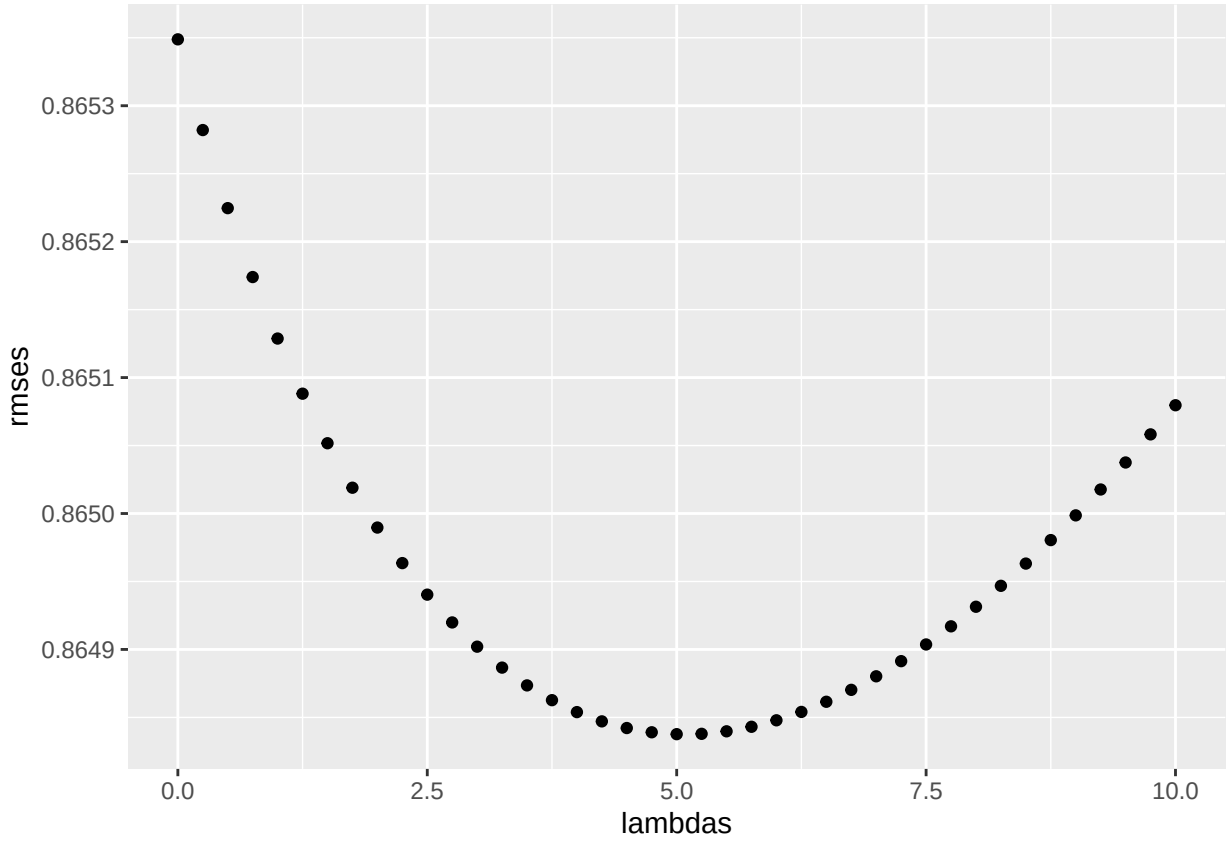
3. Results

The results section of the Report.R script is commented as "Results".

The lambda parameter that minimises RMSE is

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

This can be visualised here:



4. Conclusion

The conclusion of this report is that the improved Naive Bayes model that takes into account movie, user and decade produces an RMSE below 0.8649, as seen in the table below.

method	RMSE
Regularized Movie, User and Time Effect Model	0.8648377

This report has limitations, insofar as it has not explored the use of multiple predictive models or ensemble of models, which could yield better overall results. Future work could also include rounding of half-ratings, given that integers are predominant across the full data set.