MovieLens

Mauro Berlanda

4/5/2020

Introduction/Overview/Executive Summary

The goal of this project is to develop an algorithm to predict movie ratings.

This analysis is performed on (MovieLens) dataset. It contains 10 millions ratings and 100 000 tags applications applied to 10 000 movies by 72 000 users.

After downloading the entire dataset, two dataframes are created:

- `edx` which is used to perform the analysis and develop the algorithm to predict ratings
- `validation` which contains the true values of the predictions

```
dim(edx)
```

```
## [1] 9000055       6
```

```
dim(validation)
```

```
## [1] 999999      6
```

```
colnames(edx)
```

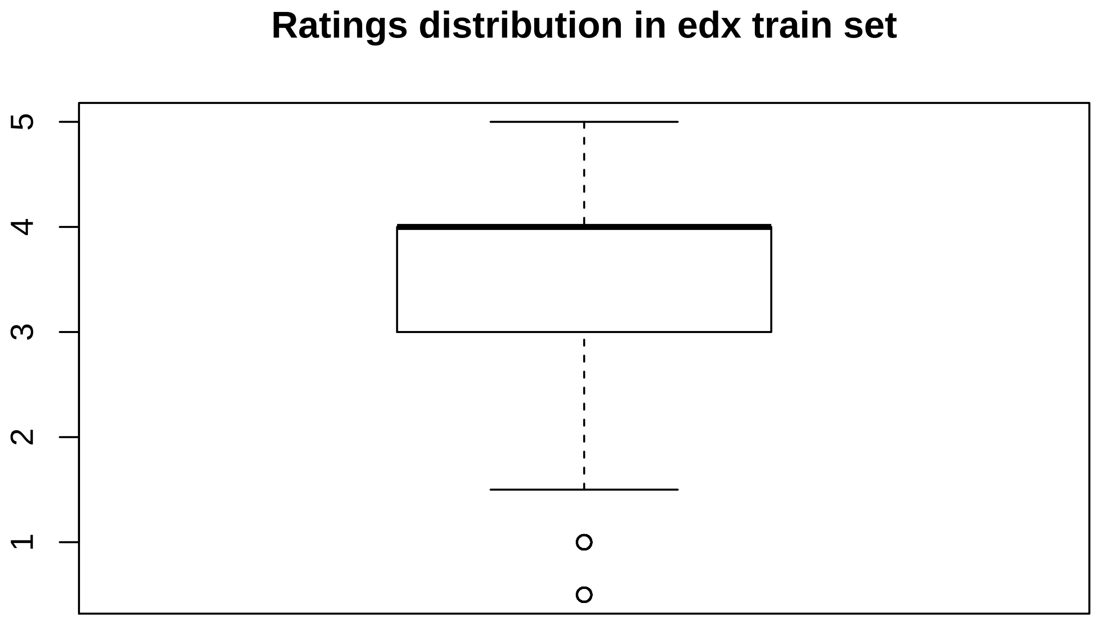
```
## [1] "userId"    "movieId"   "rating"    "timestamp" "title"     "genres"
```

The key steps required to identify an optimal solution are:

- todo
- create a train and a validation set
- explore the content of the train set
- another step

Method/Analysis

```
boxplot(edx$rating)
title("Ratings distribution in edx train set")
```



```r
summary(edx$rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.500   3.000   4.000   3.512   4.000   5.000
```

Results

Conclusion