Report on Rating Predictions for Movielens

Oliver Brueckner

2019-12-19

Introduction

This is a report on rating predictions for the movielense data set. After some data exploration we will follow some approaches to reach our goal: Make predictions for the ratings in the validation set with an RMSE <= 0.8649.

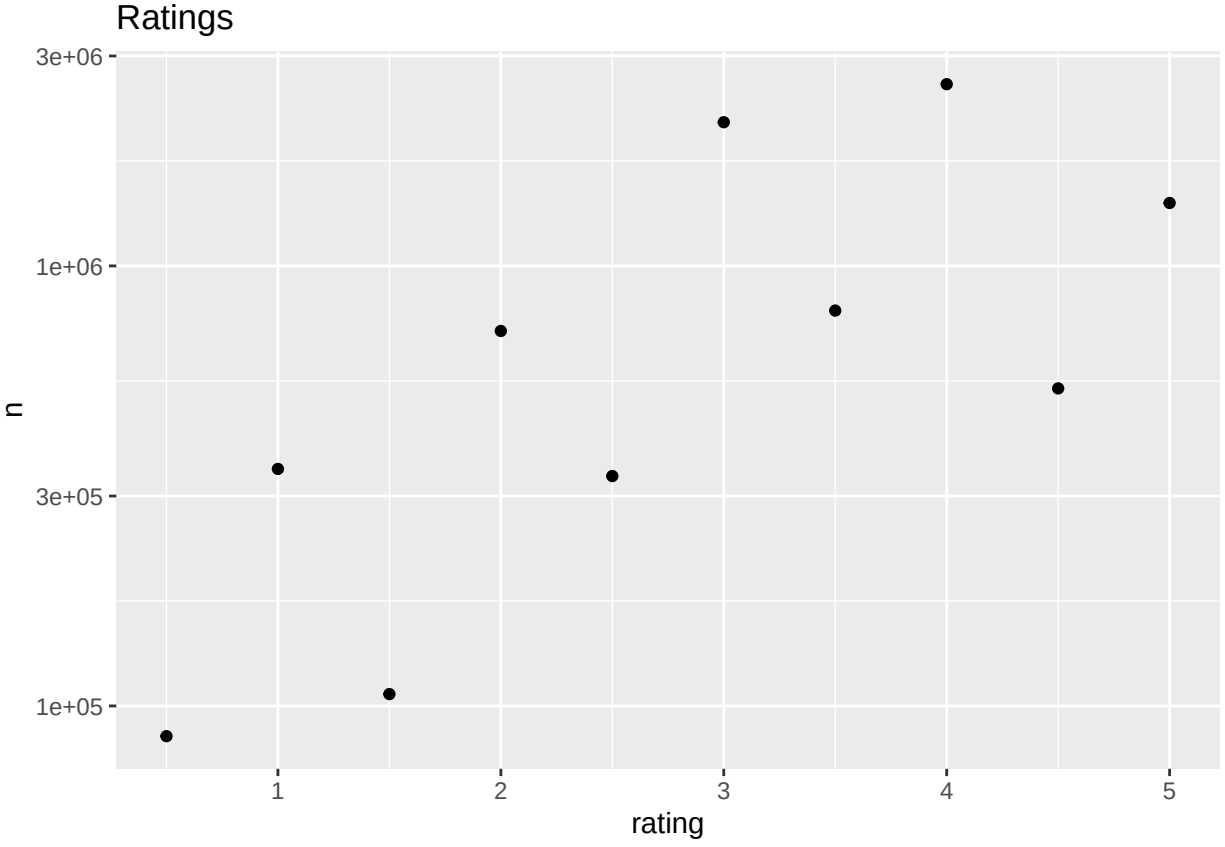
Analysis: Inspecting the Data

The data was loaded with the code given in the course. A first view on the data shows that there is a huge dataset with more than 9 million ratings on movies. These ratings come from 69,878 users and are given to 10,677 movies. These numbers show, that not every movie is rated by every user.

```
##   number_of_movies number_of_users number_of_ratings
## 1            10677           69878           9000055
```

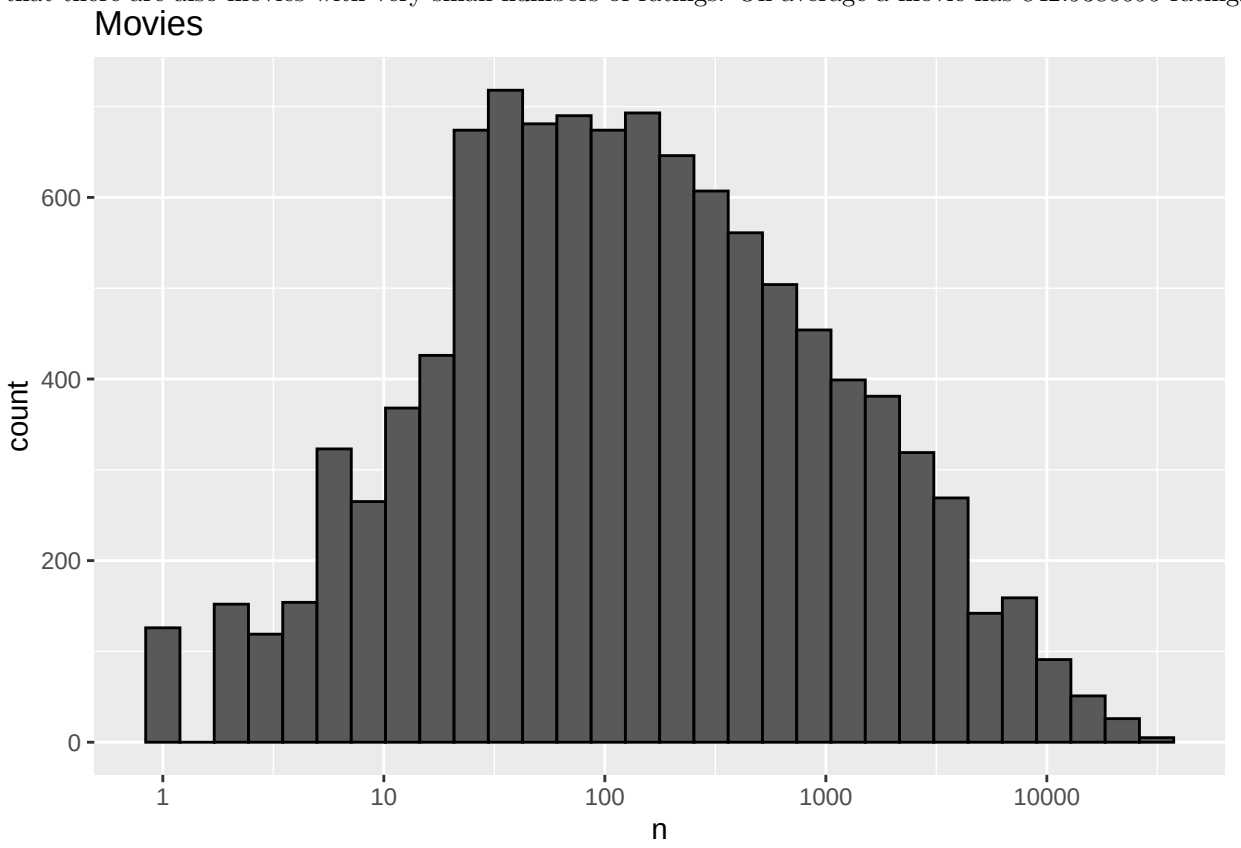
What is the Distribution of Ratings?

If we have a closer look to the distribution of the Ratings we see, that they tend to be .0 instead of .5 The average rating is about 3.51 with a standard deviation of 1.06.



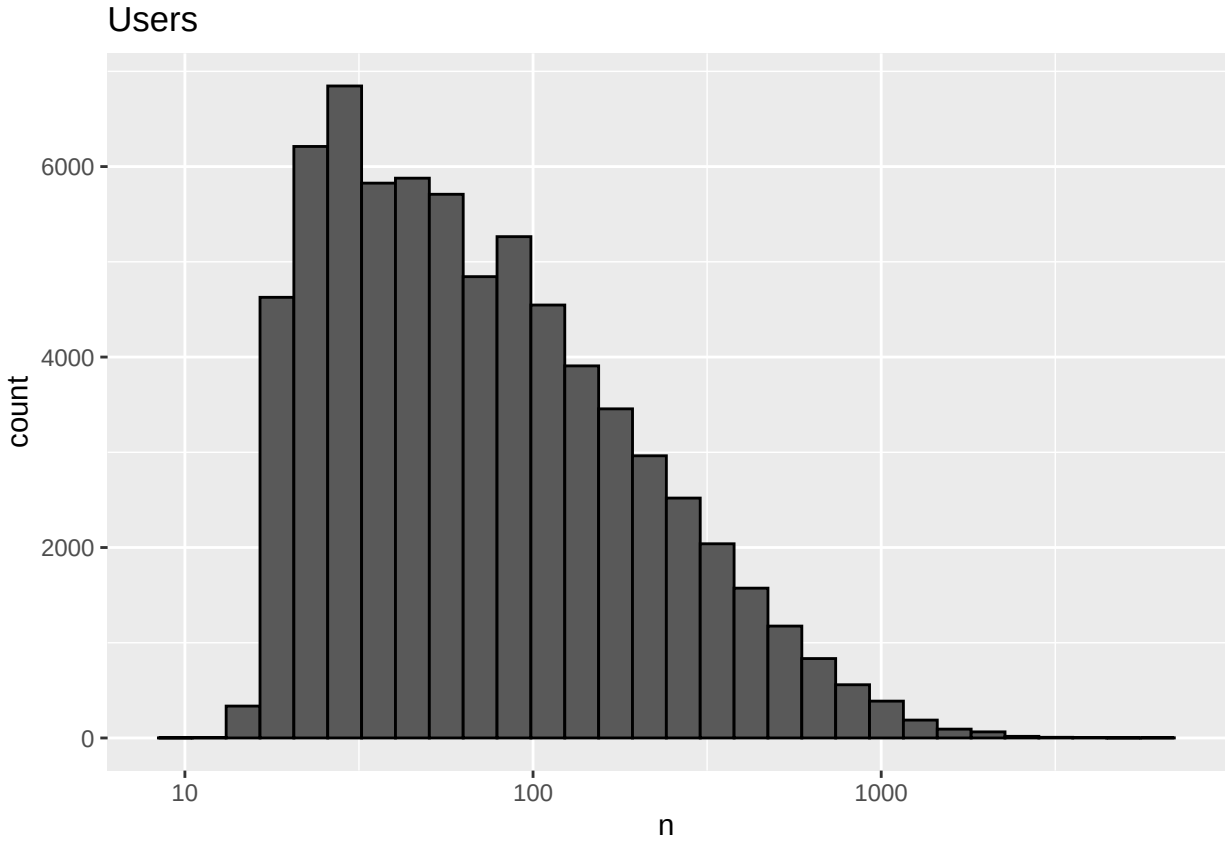
What ist the Distribution of Ratings per Movie?

If we have a look on the distinct movies we see, that there are movies with very high numbers of ratings and that there are also movies with very small numbers of ratings. On average a movie has 842.9385595 ratings.



What ist the Distribution of Ratings per User?

If we have a look on users we see, that there are users with very high numbers of ratings and that there are many users with very small numbers of ratings. On average a user has 129 ratings



Method

Our Inspection of the data shows, that the movie and the user could have an influence on our predictions. To verify this, we first try a naive approach, where we guess allways the mean rating mu_hat. Here are the results of this naive approach:

```
mu_hat
```

```
## [1] 3.512465
```

```
##             method     RMSE
## 1 Just the average 1.061202
```

As we see, this approach does not fit our goal. The RMSE is 1.06.

Using the Movie Signals to improve the guess

In the next step, we start building a linear model. The first parameter that influences our guess is the movie rating. For this we push our estimatet rating in the direction of the mean rating of the movie.

If we do this we get the following results:

```
##               method      RMSE
## 2 Movie Effect Model 0.9439087
```

As we see, the RMSE improved, but still does not fit our goal.

Using the Movie Signals and User Signals to improve the guess

The next step is adding the user signals to our model. If we do this we get the following results:

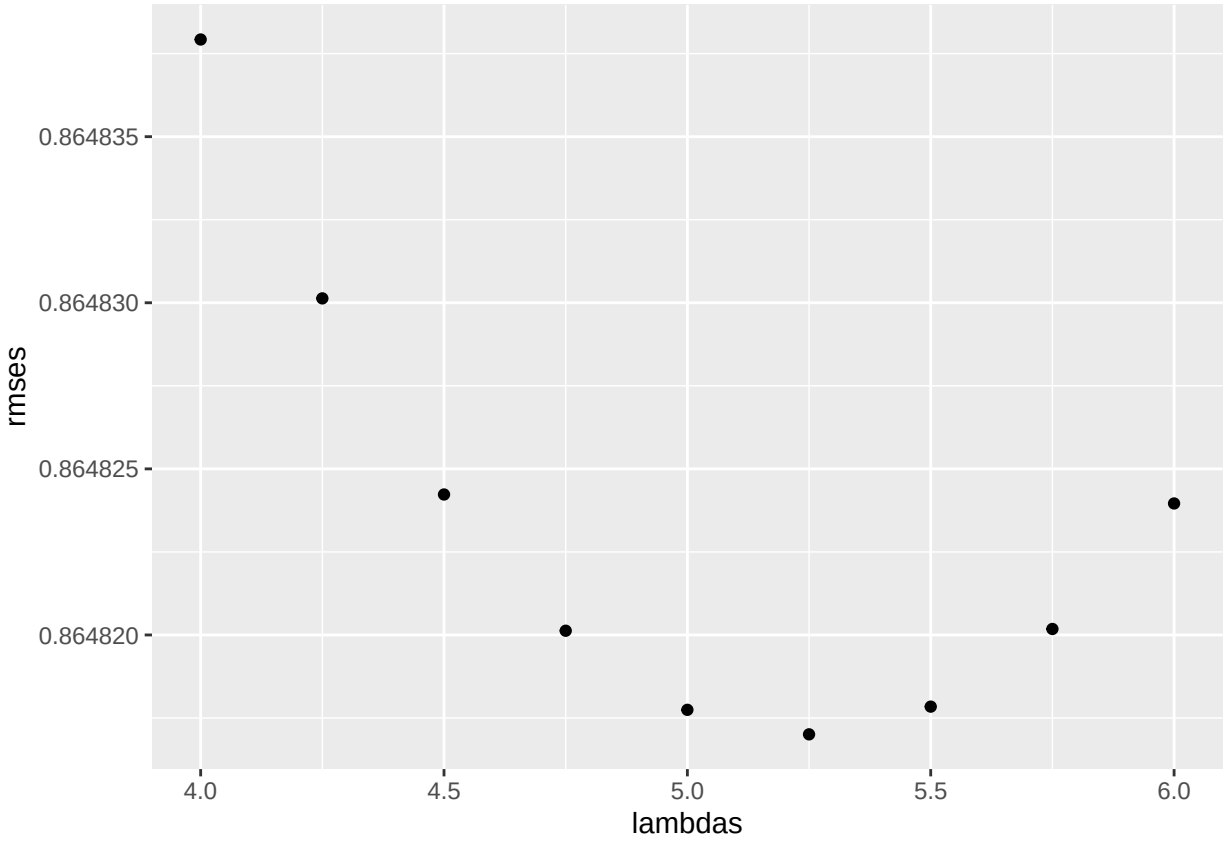
```
##                         method      RMSE
## 3 Movie + User Effects Model 0.8653488
```

As we see, the RMSE improved, but still does not fit our goal.

Using regularization to avoid large impact of small samples

In the next step we regularize the parameters of our model. This avoids large impact of small sample sizes. Regularization uses a parameter Lambda, that could be optimized. We have done that and found that Lambda=5.25 is optimal for our estimation.

```
qplot(lambdas, rmses)
```



If we use this parameter this we get the following results:

```
##                                   method     RMSE
## 4 Regularized Movie + User Effect Model 0.864817
```

As we see, the RMSE improved and fits our goal.

Results

The followin table shows the RMSE values for the approaches. We see that the regularized use of the movie and user signals is useful to make a good prediction.

```
##                                    method      RMSE
## 1                      Just the average 1.0612018
## 2                    Movie Effect Model 0.9439087
## 3            Movie + User Effects Model 0.8653488
## 4 Regularized Movie + User Effect Model 0.8648170
```

Conclusion

We see that we can make good estimations based on the movie and user signals when we avoid the effects of small sample sizes through regularization. To make our predictions more accurate the next steps could be:

- Using the genres as additional signal
- try to find similar movies and users to enhance the prediction
- add more predictors like: actors, directors or studios