Predicting user ratings by movie's reviews

Tatiana Smirnova

January 2023

Abstract

This document is a report for "Predicting user ratings by movie's reviews" course project. A link to project code: `https://github.com/smirnovata/nlp`.

1 Introduction

The main target is to build a model to predict user ratings based on the user's reviews. This is important for writing a plausible reviews. The result can show correlation between word usage and estimated ranking.

This task is predict a number and text is as input, this can be considered a classic problem for machine learning(regression).

1.1 Team

Course project was made by **Tatiana Smirnova**

2 Related Work

One article can be found that used same dataset for the All-Russian Olympiad of schoolchildren([С. Г. Григорьев, 2022]).

No solutions were found specifically on project topic, but a similar tasks were on several articles([Oluwatofunmi Adetunji, 2020], [Tran, 2022] or [Yichen Yang, 2019]). They describe user ratings prediction with another attributes.

3 Model Description

The task is to apply and compare several well-known algorithms, as well as their ensembles on the same dataset, therefore, the exact indication of the internal formulas does not matter.

The first group of methods is linear regression, for which tfidf encoding is provided.

For neural networks, tokenization is used first, and then an embedding layer in the models themselves. The ensembles look like a collection of the best models of each type, as well as one model combining the best from the past.

4 Dataset

The dataset is built using the «Kinopoisk's movies reviews» Kaggle dataset. and rating information from Kinopoisk site.

Collection of reviews for movies from kinopoisk.ru contains 131583 reviews. Every review in separate file in corresponding folder.

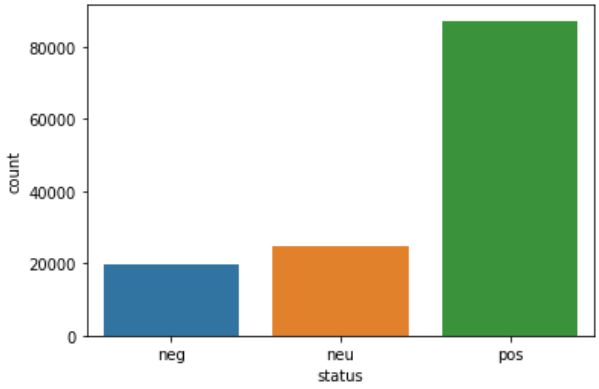

Figure 1: Number of Positive, Negative and Neutral reviews

Each review named in a way that 1st part is movie ID at kinopoisk.ru and 2nd part is review number. This ID is used to get the movie's rating.

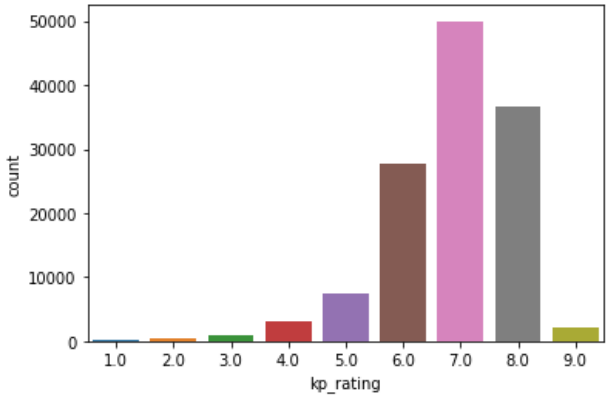

Figure 2: Number of rounded movie ratings

5 Experiments

5.1 Metrics

Root Mean Squared Error(RMSE) was used in the tasks.

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(\frac{d_i - f_i}{n}\right)^2}$$

where d_i - value of the variable calculated, f_i - value of the variable known, n- number of variables.

5.2 Experiment Setup

Baseline setup: standart linear regression from sklearn.

Hypothesis setup: Simple net: first layer - 1024 ('relu'), second layer - 512('relu'), dropout - 0.2, loss - 'mse', optimizer - 'adam'

Conventional nn: first layer conventional - 1024x7 ('relu'), second layer conventional - 1024x5('relu'), third layer conventional - 1024x3('relu'), fourth layer - 512('relu'), dropout - 0.2, loss - 'mse', optimizer - 'adam'

Ensemble of simple nn: first layer - 100 ('relu'), dropout - 0.2, loss - 'mse', optimizer - 'adam'. And 10 nn with 1 epoch training.

5.3 Baselines

In our case using TFIDF with linear regression as baselines.

6 Results

RMSE best result:

1. Baseline(linear regression) have value - 0.905
2. Simple net - 0.949
3. Conventional net - 0.909
4. Ensemble of simple net - 0.903

The best result in RMSE - 0.903.

7 Conclusion

As a result, we collected a dataset, made a markup for it and developed a model showing the best results compared to other models.

References

[Oluwatofunmi Adetunji, 2020] Oluwatofunmi Adetunji, Mamudu Hadiza, O. N. (2020). Design of a movie review rating prediction (mr2p) algorithm. *International Journal of Scientific Research in Computer Science Engineering and Information Technology*. URL: https://www.researchgate.net/publication/343686586_Design_of_a_Movie_Review_Rating_Prediction_MR2P_Algorithm.

[Tran, 2022] Tran, K. (2022). Predict movie ratings with user-based collaborative filtering. *Towards Data Science*. URL: https://towardsdatascience.com/predict-movie-ratings-with-user-based-collaborative-filtering-392304b988af.

[Yichen Yang, 2019] Yichen Yang, Ruoyun Ma, M. H. C. (2019). Predicting movie ratings with multimodal data. *CS229: Machine Learning*. URL: https://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/26260680.pdf.

[С. Г. Григорьев, 2022] С. Г. Григорьев, И. А. Калинин, (2022). Система заданий для первой всероссийской олимпиады школьников по искусственному интеллекту. *Журнал. Информатика и образование*. URL: https://info.infojournal.ru/jour/article/view/831.